Trabajo Práctico N° 2

Procesamiento del lenguaje natural

Tecnicatura Universitaria en Inteligencia Artificial

FCEIA - UNR

2do año

Integrantes

Fernández, Florencia

1. Introducción

Este informe busca mostrar y comentar la resolución de los ejercicios planteados mediante aplicación de las unidades teóricas vistas para resolver problemas relacionados con la conversación del usuario con un chatbot experto y propuestas a tratar problemáticas con sistemas multiagentes. Las actividades incluyen limpieza de datos, modularización del código, entendimiento del contexto, enfoque específico por temática, recolección de fuentes de datos adecuadas para la tarea e iteraciones del proceso en busca del mejor resultado.

1.1 - Base de datos de vectores

En esta fuente de datos externa, se tendrá información y datos relacionados con la enseñanza de vocabulario y técnicas para aprender costura, diseño y patronaje.

1.1.1 - Función para obtener el texto de los archivos fuente

Esta función guarda los textos redactados en los documentos en formato string en un diccionario cuyas claves son los nombres de los archivos y el valor correspondiente es el contenido del archivo.

1.1.2 - Se separan los textos en chunks

Se utiliza NLTKTextSplitter que por defecto separa por doble salto de línea. Con esto se busca no cortar una oración sin que haya terminado como podría pasar si se usa un chunk con cantidad de caracteres fijo. De esta manera, se busca capturar mejor el contenido semántico del texto.

1.1.3 - Preprocesamiento de los datos

Se han definido distintas funciones que se utilizarán para preprocesar el texto que será vectorizado luego, buscando con las mismas obtener un texto más enfocado en los significados intrínsecos de las oraciones y el contexto y con menos ruido.

- Función para remover acentos
  La eliminación de acentos simplifica el texto y evita posibles problemas de codificación que podrían surgir con caracteres acentuados.

- Función para reemplazar palabras
  Se reemplaza una abreviatura muy común en el ámbito de la costura por su palabra completa para que sea semánticamente más destacable. Además, se extienden palabras que tienen sinónimos utilizados en distintos lugares del mundo en español para que formen parte de los embeddings que brindarán información al modelo.

- Función para remover los mails y urls del texto
  Se remueven porque se considera que no aportan a la enseñanza de técnicas y conceptos para el tema propuesto.

- Función para remover stopwords
  Las stopwords son palabras comunes como "el", "la", "y", etc., que no aportan mucha información en términos de contexto y significado. Se eliminan porque ayuda a reducir la dimensionalidad del texto y a concentrarse en las palabras clave.

- Función que engloba las anteriores para ser utilizada sobre el texto
  Para cada chunk de texto, se aplica el preprocesamiento utilizando las funciones anteriores y luego se muestra el resultado.

1.1.4 - Vectorización del texto

Se utiliza un modelo de embeddings para vectorizar el texto ya que tiene en cuenta la semántica y contexto de lo que representa. Se busca así, robustecer la representación del texto al pasarlo a vectores.
Se añade a la colección de la base de datos vectorial Chromadb los fragmentos de texto convertidos a vectores.
Luego, realiza una consulta para conocer la respuesta de la base de datos a la misma.

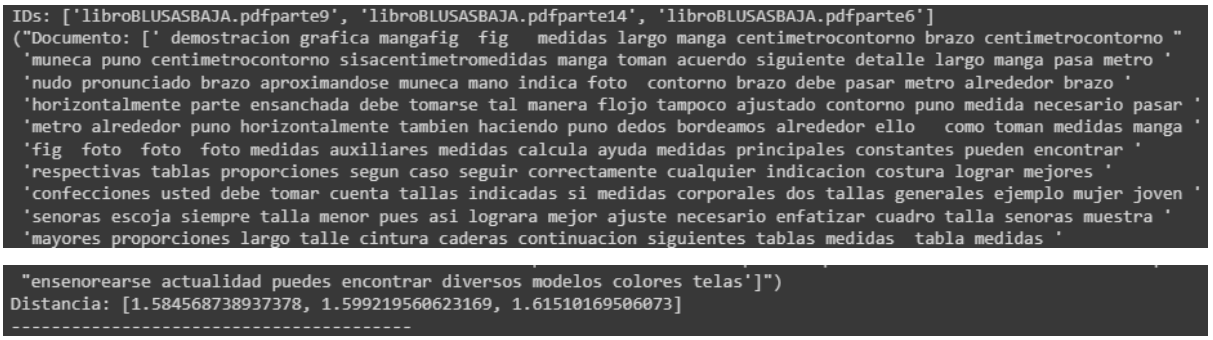


Imagen que muestra el resultado de la información proporcionada por la base de datos de vectores para la consulta "Cómo se cose una remera?"
Se responde con los 3 resultados más similares

1.2 - Base de datos de grafos

Esta fuente de datos externa será utilizada para abarcar las búsquedas de personalidades históricas relevantes en el ámbito de la costura y la moda.

La base de datos de grafos que se utilizará es dbpedia.
Se definen las distintas funciones que permitirán la interacción con la base de datos de grafos. En órden de aparición:

- Función para realizar la consulta a la base de datos de grafos
- Función para obtener el nombre de la personalidad por la que se consulta:
  Esta función analiza la query en busca de NER (entidades nombradas) y las extrae de la misma. Retornando una lista de los nombres hallados.
- Función que arma la la consulta en lenguaje SPARQL para que pueda ser entendida por la base de datos de grafos y luego, realiza la consulta a la misma.

Se muestra un ejemplo de uso y los resultados obtenidos

```
Entidad: Giorgio Armani, Etiqueta: PER, Explicación: Named person or family.
Entidad: Carolina Herrera, Etiqueta: PER, Explicación: Named person or family.
URI del diseñador Giorgio Armani: http://es.dbpedia.org/resource/Giorgio_Armani
Información sobre Giorgio Armani en DBpedia:
Giorgio Armani (Piacenza, 11 de julio de 1934) es un diseñador de moda y empresario de origen armenio-italiano. Principalmente
URI del diseñador Carolina Herrera: http://dbpedia.org/resource/Carolina_Herrera
Información sobre Carolina Herrera en DBpedia:
María Carolina Josefina Pacanins Niño (Caracas; 8 de enero de 1939), mejor conocida como Carolina Herrera, es una diseñadora de
```

Imagen que muestra el resultado de la información proporcionada por la base de datos de grafos dbpedia para las consultas `"Busca información sobre la diseñadora de moda Giorgio Armani y Carolina Herrera."`
Se muestra el mejor resultados hallado

1.3 - Datos en formato tabular

En esta fuente de datos externa, se cuenta con los datos numéricos que corresponden a las medidas de talles para distintas secciones del cuerpo. Será consultada en caso de que se solicite información relacionada a obtener mediciones de referencia.
Se leen los archivos tabulares y se convierten a un string que conserva su estructura tabular.

```
TALLAS_DE_NIÑOS_AS_EN_CENTIMETROS
              EDAD     2     4    6    8   10   12   14
 CONTORNO DE CUELLO 24,5  26,5 28,5 30,5 32,5 34,5 36,5
 CONTORNO DE PECHO    54    58   62   66   72   76   80
CONTORNO DE CINTURA   50    53   56   59   61   63   65
CONTORNO DE CADERA    58    62   66   70   74   78   82
    LARGO DE TALLE  18,5  21,5 24,5 27,5 29,5 31,5   33
     LARGO A CADERA   10  11,5   13 14,5   16 17,5   19
    LARGO A RODILLA 32,5 36,75 39,5   43 46,5   50 53,5
       LARGO HOMBRO  7,2     8  8,7  9,3  9,9 10,5   11
     LARGO DE BRAZO   30    34   38   42   46   50   54
     LARGO PANTALON   56    62   68   74   80   86   92
           ESTATURA   92   102  112  122  132  142  152
TABLA_DE_MEDIDAS_MASCULINAS_EN_CENTIMETROS
 TOMAR MEDIDAS EN      SECTOR  TALLA S (38) HOMBRE TALLA S (40) HOMBRE  TALLA M (42) HOMBRE TALLA M (44) HO
         SECTOR ABREVIATURA                   38                  40                  42
           PECHO           P                  43                  45                  47
TALLE DE ESPALDA          TE                  41                  42                  43
ANCHO DE ESPALDA          AE                  42                  43                  44
        COSTADO           CO                  19                19,5                  20                  2
        CINTURA           CT                  33                  35                  37
         CADERA           CD                  39                  41                  43
         CUELLO           CU                  36                  37                  38
  LARGO DE MANGA          LM                  54                  55                  56
TABLA_DE_MEDIDAS_FEMENINAS_EN_CENTIMETROS
    TOMAR MEDIDAS EN      SECTOR TALLA S TALLA S.1 TALLA M TALLA M.1 TALLA L TALLA L.1 TALLA XL TALLA XL.1
            SECTOR ABREVIATURA      38       40       42       44       46       48       50         52
              BUSTO             B      40       43       45       48       50       53       55         58
   TALLE DE ESPALDA          TE      41      41,5      42      42,5      43      43,5      44        44,5
   ANCHO DE ESPALDA          AE      34       36       37       39       40       42       43         46
            COSTADO          CO      19      19,1     19,3     19,4     19,5     19,7     19,8         20
      ALTO DE BUSTO          AB      24       25       26       27       28       29       30         31
      BUSTO A BUSTO          BB    18,5       19     19,5       20     20,5       21     21,5         22
            CINTURA          CT      32       34       36       38       40       43       45         47
             CADERA          CD      41       44       47       50       52       55       57         60
CINTURA A LA CADERA          CC      16      16,5       17       18     18,5       19       20         20
            CUELLO           CU      34       35       36       37       38       39       40         41
     LARGO DE MANGA          LM      54       55       56       57       58       59       60         61
```

Imagen que muestra el resultado de la información tabular. Se conforma de 3 tablas de medidas de distintas secciones del cuerpo para distintos talles de niño, mujer y hombre

1.4 - Elección de la fuente externa para la consulta

A continuación, se requiere definir un criterio para determinar la fuente que resulte más apta para proveer los datos de contexto que permiten una respuesta más precisa y exacta de la consulta.
Se ha utilizado un modelo liviano para realizar esta elección.
Inicialmente, se han definido ejemplos de consultas que coinciden con cada fuente externa en la variable few_shots_examples.
Luego, se ha tomado la consulta del usuario, y se ha calculado la similaridad de dicha consulta con las consultas de ejemplo de las distintas fuentes. La fuente con la que se ha encontrado mayor similaridad, es la seleccionada como la más apta. Para ser seleccionada, además, debe superar un umbral de similaridad. Con esto se busca que consultas muy alejadas de las propuestas como patrones, no sean consideradas a ser contextualizadas con las fuentes externas.

```
consulta_usuario1 = "Cómo coser un pantalón?"
consulta_usuario2 = "Cuáles son las medidas de talles de pantalón para mujer?"
consulta_usuario3 = "Quién fue Gianni Versace?"
fuente_seleccionada1 = few_shot_learning(consulta_usuario1, few_shot_examples)
fuente_seleccionada2 = few_shot_learning(consulta_usuario2, few_shot_examples)
fuente_seleccionada3 = few_shot_learning(consulta_usuario3, few_shot_examples)
print(f"La fuente de datos seleccionada para la consulta {consulta_usuario1} es: {fuente_seleccionada1} y se esperaba que fuera vectores")
if fuente_seleccionada1 == 'vectores':
  print('Se ha seleccionado la fuente esperada')
else:
  print('No se ha seleccionado la fuente esperada')
print('\n')
print(f"La fuente de datos seleccionada para la consulta {consulta_usuario2} es: {fuente_seleccionada2} y se esperaba que fuera csv")
if fuente_seleccionada2 == 'csv':
  print('Se ha seleccionado la fuente esperada')
else:
  print('No se ha seleccionado la fuente esperada')
print('\n')
print(f"La fuente de datos seleccionada para la consulta {consulta_usuario3} es: {fuente_seleccionada3} y se esperaba que fuera grafo")
if fuente_seleccionada3 == 'grafo':
  print('Se ha seleccionado la fuente esperada')
else:
  print('No se ha seleccionado la fuente esperada')
print('\n')
```

```
La fuente de datos seleccionada para la consulta Cómo coser un pantalón? es: vectores y se esperaba que fuera vectores
Se ha seleccionado la fuente esperada


La fuente de datos seleccionada para la consulta Cuáles son las medidas de talles de pantalón para mujer? es: csv y se esperaba que fuera csv
Se ha seleccionado la fuente esperada


La fuente de datos seleccionada para la consulta Quién fue Gianni Versace? es: grafo y se esperaba que fuera grafo
Se ha seleccionado la fuente esperada
```

Imagen que muestra el resultado de la fuente externa elegida como aquella apta para que la información que proporciona para la consulta sea añadida al contexto de la consulta que se solicita al LLM

1.5 - RAG (Generación Aumentada por Recuperación)

Los LLMs de última generación son entrenados con grandes cantidades de datos para lograr un amplio espectro de conocimiento general almacenado en los pesos de la red neuronal (memoria paramétrica). Sin embargo, al solicitar a un LLM que genere una respuesta que requiere conocimientos que no estaban incluidos en sus datos de entrenamiento, como información más reciente, propietaria o específica de un dominio, puede llevar a inexactitudes fácticas, o bien puede no conocer la respuesta.

La Generación Aumentada por Recuperación (RAG) es el concepto de proporcionar a los Modelos de Lenguaje de Gran Escala (LLMs) información adicional proveniente de una fuente de conocimiento externa. Esto les permite generar respuestas más precisas y contextuales.
En este caso, se ha proporcionado al LLM zephyr-7b-beta, información de contexto adicional para la temática de 'Corte y confección de prendas, patronaje y diseño'.
Una vez definidas las funciones y fuentes precisadas para dar contexto al modelo LLM se han definido funciones pertinentes para implementar el RAG y realizar las consultas al mismo.

- Función que tiene las instrucciónes de formación del prompt que se envía al modelo para realizar la consulta
- Función que adiciona a la consulta el contexto brindado por la fuente externa y los transforma en el prompt específico de dicha consulta para procesar por el modelo.
- Función que analiza la query y devuelve la respuesta encontrada haciendo uso de las funciones anteriores
- Función que realiza lo mismo que la función global anterior de manera interactiva con el usuario

```
Escriba su consulta: Receta de alfajores
Selected context source: No se ha detectado que la consulta pueda ser respondida por las distinta
Pregunta: Receta de alfajores
Respuesta:
Ingredientes:
- 250 gramos de harina
- 150 gramos de azúcar
- 100 gramos de mantequilla
- 100 gramos de miel
- 100 gramos de semillas de jazmínez
- 1 cucharada de canela molida
- 1 cucharada de vainilla extracto
- 150 ml de leche
- Azúcar para espolvorear

Instrucciones:
1. En un tazón, calienta la leche hasta el punto de ebullición.
2. Retira de la fuente de calor y agrega la miel, mezcla hasta que se disuelva.
3. Deja enfriar la mezcla hasta que tenga una temperatura de 40-45 grados Celsius.
4. En un recipiente grande, mezcle la harina, el azúcar, la mantequilla, la semilla de jazmínez,
5. Agrega la mezcla de leche y miel a la mezcla de harina y mezcle hasta que se forme una masa su
6. Deja reposar la masa durante 30 minutos en el recipiente.
7. Forma bolas pequeñas con la masa y coloca en una bandeja para hornear.
8. Hornea en el horno a 160 grados Celsius durante 20-25 minutos, o hasta que se doren ligerament
9. Quita de la bandeja y deja enfriar.
10. Espolvorea con azúcar para decorar.
```

```
Esperamos que disfrutes de estos deliciosos alfajores!
----------------------------------------------------------
Desea realizar otra consulta? Si desea realizar otra consulta escriba si: si
Escriba su consulta: Quién fue Giorgio Armani?
Selected context source: grafo
Entidad: Giorgio Armani, Etiqueta: PER, Explicación: Named person or family.
Pregunta: Quién fue Giorgio Armani?
Respuesta:
Giorgio Armani es un diseñador de moda y empresario italiano de origen armenio, conocido principalmente p
----------------------------------------------------------
Desea realizar otra consulta? Si desea realizar otra consulta escriba si: si
Escriba su consulta: Cómo aprender costura para principiantes?
Selected context source: vectores
Pregunta: Cómo aprender costura para principiantes?
Respuesta:
Para aprender costura para principiantas, es necesario seguir un proceso de capacitación que se centra en

Primero, es importante conocer las medidas y normas de seguridad implementadas en el taller de costura. E

Seguidamente, es importante entender los diferentes tipos de maquinaria y herramientas utilizados en la c

Es importante también conocer las diferentes medidas y tipos de puntadas, como la puntada simple, la punt

Además, es importante conocer los diferentes tipos de puntadas y costuras, como la cadeneta simple, la ca

Para aprender costura, es necesario practicar diferentes ejercicios y tareas, como el corte de la tela, l

Además, es importante conocer las diferentes medidas y tipos de puntadas, como la puntada simple, la punt

Para mejorar la calidad y la perfección de la costura, es necesario tomar medidas precisas y tener un con

En resumen, para aprender costura para principiantas, es necesario seguir un proceso de capacitación que
----------------------------------------------------------
Desea realizar otra consulta? Si desea realizar otra consulta escriba si: si
Escriba su consulta: Cuáles son los talles de niños?
Selected context source: csv
Pregunta: Cuáles son los talles de niños?
Respuesta:
Los talles de niños se encuentran en la información de contexto siguiente:
```

```
---------------------
TALLAS DE NIÑOS AS-EN CENTÍMETROS
              EDAD     2     4    6    8   10   12   14
 CONTORNO DE CUELLO 24,5  26,5 28,5 30,5 32,5 34,5 36,5
 CONTORNO DE PECHO    54    58   62   66   72   76   80
CONTORNO DE CINTURA   50    53   56   59   61   63   65
CONTORNO DE CADERA    58    62   66   70   74   78   82
    LARGO DE TALLE  18,5  21,5 24,5 27,5 29,5 31,5   33
     LARGO A CADERA   10  11,5   13 14,5   16 17,5   19
    LARGO A RODILLA 32,5 36,75 39,5   43 46,5   50 53,5
       LARGO HOMBRO  7,2     8  8,7  9,3  9,9 10,5   11
     LARGO DE BRAZO   30    34   38   42   46   50   54
     LARGO PANTALON   56    62   68   74   80   86   92
           ESTATURA   92   102  112  122  132  142  152
---------------------

Así, todas las tablas de talles están presentes en la información de contexto.
----------------------------------------------------------
Desea realizar otra consulta? Si desea realizar otra consulta escriba si: no
```

Imagen que muestra las respuestas obtenidas al aplicar la técnica RAG para las consultas "Receta de alfajores", "Quién fue Giorgio Armani?", "Cómo aprender costura para principiantes?", "Cuáles son los talles de niños?"

Se observa que asigna la consulta a la fuente esperada y, que cuando no encuentra que corresponda asignar una de las tres fuentes para dicha consulta, responde sin añadir contexto específico con lo que devuelve el LLM para la consulta.

2. Ejercicio 1 - RAG

Crear un chatbot experto en un tema a elección, usando la técnica RAG (Retrieval Augmented Generation). Como fuentes de conocimiento se utilizarán al menos las siguientes fuentes:

- Documentos de texto
- Datos numéricos en formato tabular (por ej., Dataframes, CSV, sqlite, etc.)
- Base de datos de grafos (Online o local)

El sistema debe poder llevar a cabo una conversación en lenguaje español. El usuario podrá hacer preguntas, que el chatbot intentará responder a partir de datos de algunas de sus fuentes. El asistente debe poder clasificar las preguntas, para saber qué fuentes de datos utilizar como contexto para generar una respuesta.

Requerimientos generales:

- Realizar todo el proyecto en un entorno Google Colab
- El conjunto de datos debe tener al menos 100 páginas de texto y un mínimo de 3 documentos.
- Realizar split de textos usando Langchain (RecursiveTextSearch, u otros métodos disponibles). Limpiar el texto según sea conveniente.
- Realizar los embeddings que permitan vectorizar el texto y almacenarlo en una base de datos ChromaDB
- Los modelos de embeddings y LLM para generación de texto son a elección

3. Ejercicio 2 - Agentes

Realice una investigación respecto al estado del arte de las aplicaciones actuales de agentes inteligentes usando modelos LLM libres.

Plantee una problemática a solucionar con un sistema multiagente. Defina cada uno de los agentes involucrados en la tarea.

Realice un informe con los resultados de la investigación y con el esquema del sistema multiagente, no olvide incluir fuentes de información.

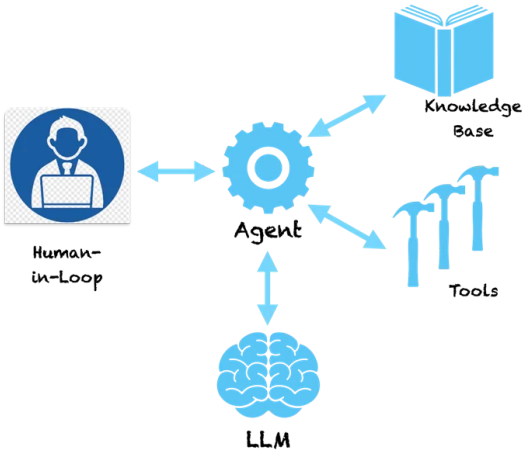
Opcional: Resolución con código de dicho escenario.

3.1 Introducción

Los Agentes Inteligentes, impulsados por modelos de lenguaje de aprendizaje profundo (LLM), han experimentado un crecimiento significativo en diversas aplicaciones. Estos agentes son capaces de comprender y generar texto de manera inteligente, lo que abre la puerta a una amplia gama de aplicaciones en el procesamiento del lenguaje natural (PLN).

3.2 Investigación de aplicaciones actuales de agentes inteligentes

En inteligencia artificial, un agente inteligente es un sistema perceptivo capaz de interpretar y procesar la información que recibe de su entorno, actuando en consecuencia de acuerdo a los datos que recoge y procesa.



En el desarrollo constante de la inteligencia artificial, los agentes inteligentes cobran cada vez más relevancia en el ámbito tecnológico, y es que, su extrema utilidad en los numerosos aspectos del mundo real, hacen de esta entidad autónoma un sistema que se aplicará cada vez más, y seremos testigos de sus grandes ventajas cuando experimentemos los beneficios que aportan a la humanidad.

3.2.1 Aplicaciones de agentes inteligentes: usos y ámbitos

Las notables capacidades de los LLM han dado lugar a una gran cantidad de aplicaciones en diversas industrias y dominios. La siguiente lista está lejos de ser exhaustiva, pero aborda algunos de los casos de uso más populares y útiles detrás de los LLM.

- Traducción automática:
  El objetivo es traducir automáticamente texto o voz de un idioma a otro. Los LLM, como T5 de Google y la serie GPT de OpenAI, han logrado un rendimiento notable en tareas de traducción automática, reduciendo las barreras del idioma y facilitando la comunicación intercultural. Uno de los ejemplos más conocidos es Google Translate.

- Análisis de sentimientos:
  Implica determinar el sentimiento o la emoción expresada en un texto, como una reseña de un producto, una publicación en las redes sociales o un artículo de noticias. Los LLM pueden extraer de manera efectiva la información de sentimientos de los datos de texto, lo que permite a las empresas medir la satisfacción del cliente, monitorear la reputación de la marca y descubrir información para el desarrollo de productos y estrategias de marketing.
    - Los especialistas en marketing pueden entrenar un modelo de lenguaje grande para organizar los comentarios y las solicitudes de los clientes en grupos, o segmentar productos en categorías según las descripciones de los productos. En base a ello se pueden desprender mejoras en tareas tales

como: lanzamiento de productos, vigilancia de la competencia, vigilancia de líderes de opinión y del lobbying, vigilancia de las tendencias

- Chatbots y asistentes virtuales:
  Los avances en LLM han llevado al desarrollo de sofisticados chatbots y asistentes virtuales capaces de entablar conversaciones más naturales y conscientes del contexto. Al aprovechar la comprensión del idioma y las capacidades de generación de modelos como GPT-3, estos agentes conversacionales pueden ayudar a los usuarios en diversas tareas, como atención al cliente, programación de citas y recuperación de información, brindando una experiencia de usuario más fluida y personalizada.
    - Los minoristas y otros proveedores de servicios pueden usar grandes modelos de lenguaje para brindar mejores experiencias a los clientes a través de chatbots dinámicos y asistentes de IA.

- Resumen de texto y transcripción:
  El resumen de texto consiste en generar un resumen conciso y coherente de un texto más largo, conservando su información y significado esenciales. Los LLM se han mostrado muy prometedores en esta área, ya que permiten la generación automática de resúmenes para artículos de noticias, trabajos de investigación y otros documentos extensos. Esta capacidad puede ahorrar mucho tiempo y esfuerzo a los usuarios que buscan comprender rápidamente los puntos principales de un documento.
    - Los asesores financieros pueden resumir las llamadas de ganancias y crear transcripciones de reuniones importantes utilizando grandes modelos de lenguaje. Y las compañías de tarjetas de crédito pueden usar LLM para la detección de anomalías y el análisis de fraudes para proteger a los consumidores.

- Búsqueda y generación de contenido:
  Los LLM han demostrado una capacidad excepcional para generar texto coherente y contextualmente relevante, que puede aprovecharse para tareas de generación de contenido y paráfrasis. Las aplicaciones en este dominio incluyen la creación de contenido de redes sociales y la reformulación de oraciones para mejorar la claridad de la redacción.
    - Los equipos legales pueden usar grandes modelos de lenguaje para ayudar con la escritura y la paráfrasis legal.
    - Los motores de búsqueda pueden usar grandes modelos de lenguaje para proporcionar respuestas más directas y similares a las humanas.

- Asistencia en generación de código y programación:
  Las aplicaciones emergentes de LLM en el ámbito del desarrollo de software implican el uso de modelos como Codex de OpenAI para generar fragmentos de código u ofrecer asistencia de programación basada en descripciones de lenguaje natural. Al comprender los lenguajes y conceptos de programación, los LLM pueden ayudar a los desarrolladores a escribir código de manera más eficiente, depurar problemas e incluso aprender nuevos lenguajes de programación.

    - Los desarrolladores pueden escribir software y enseñar a los robots tareas físicas con grandes modelos de lenguaje.

- Educación e investigación:
  Las capacidades de los LLM pueden ser apalancadas en entornos educativos para crear experiencias de aprendizaje personalizadas, proporcionar comentarios instantáneos sobre las tareas y generar explicaciones o ejemplos para conceptos complejos. Además, los LLM pueden ayudar a los investigadores a revisar la literatura, resumir artículos e incluso generar borradores para trabajos de investigación.
    - Según la UNESCO, la inteligencia artificial (IA) tiene el potencial de abordar algunos de los mayores desafíos de la educación actual, innovar las prácticas de enseñanza y aprendizaje, y acelerar el progreso hacia el ODS 4 (Sustainable Development Goal, por sus siglas en inglés).
    - En investigación, por ejemplo, los investigadores de ciencias de la vida pueden entrenar grandes modelos de lenguaje para comprender proteínas, moléculas, ADN y ARN.

Open LLMs:
El siguiente link contiene tablas que muestran rápida y claramente LLM (modelos de lenguajes grandes) que tienen licencia para uso comercial (por ejemplo, Apache 2.0, MIT, OpenRAIL-M) y opciones de Open LLMs para código.
Incluye información como el modelo de lenguaje, la licencia, los papers, la fecha de lanzamiento y cómo probarlo.

3.2.2 Desafíos de los LLMs

Escalar y mantener grandes modelos de lenguaje puede ser difícil y costoso.
La construcción de un modelo básico de lenguaje extenso a menudo requiere meses de tiempo de capacitación y millones de dólares.
Y debido a que los LLM requieren una cantidad significativa de datos de capacitación, los desarrolladores y las empresas pueden encontrar un desafío para acceder a conjuntos de datos lo suficientemente grandes.
Debido a la escala de los grandes modelos de lenguaje, implementarlos requiere experiencia técnica, incluida una sólida comprensión del aprendizaje profundo, los modelos de transformadores y el software y el hardware distribuidos.
Muchos líderes en tecnología están trabajando para avanzar en el desarrollo y crear recursos que puedan ampliar el acceso a modelos de lenguajes grandes, lo que permite que los consumidores y las empresas de todos los tamaños obtengan sus beneficios.
Además, no se debe ignorar, abordar las consideraciones éticas y los desafíos asociados con los modelos de lenguaje grande es un aspecto crucial de IA responsable desarrollo. Al reconocer y abordar de manera proactiva los posibles sesgos, las preocupaciones sobre la privacidad, los impactos ambientales y otros dilemas éticos, los investigadores, desarrolladores y legisladores pueden allanar el camino para un futuro impulsado por la IA más equitativo, seguro y sostenible. Este esfuerzo de colaboración puede garantizar que los LLM continúen revolucionando las industrias y mejorando vidas, manteniendo los más altos estándares de responsabilidad ética.

3.2.3 ¿Cuál es el futuro de los LLMs?

La introducción de modelos de lenguaje de gran tamaño, como ChatGPT, Claude 2 y Llama 2, que pueden responder preguntas y generar texto, apunta a interesantes posibilidades en el futuro. De forma lenta pero segura, los LLM están logrando un rendimiento similar al humano. El éxito inmediato de estos LLM demuestra un gran interés en los LLM de tipo robótico que emulan y, en algunos contextos, superan al cerebro humano. A continuación, se mencionan algunas reflexiones sobre el futuro de los LLM:

- Mayores capacidades:
  Por impresionantes que sean, el nivel tecnológico actual no es perfecto y los LLM no son infalibles. Sin embargo, las versiones más recientes mejorarán la precisión y las capacidades a medida que los desarrolladores aprendan a mejorar su rendimiento y, al mismo tiempo, reducir los sesgos y eliminar las respuestas incorrectas.

- Entrenamiento audiovisual:
  Si bien los desarrolladores entrenan a la mayoría de los LLM con texto, algunos han empezado a entrenar modelos con entrada de video y audio. Este tipo de entrenamiento debería conducir a un desarrollo de modelos más rápido y abrir nuevas posibilidades en términos de uso de LLM para vehículos autónomos.

- Transformación del lugar de trabajo:
  Los LLM son un factor disruptivo que cambiará el lugar de trabajo. Es probable que los LLM reduzcan las tareas monótonas y repetitivas de la misma manera que lo hicieron los robots con las tareas de fabricación repetitivas. Las posibilidades incluyen tareas administrativas repetitivas, chatbots de servicio al cliente y redacción automatizada y simple de textos publicitarios.

- IA conversacional:
  Sin duda, los LLM mejorarán el rendimiento de los asistentes virtuales automatizados como Alexa, Google Assistant y Siri. Podrán interpretar mejor la intención del usuario y responder a comandos sofisticados.

3.3 Problemática a solucionar con un sistema multiagente

Dada la investigación realizada anteriormente y, recordando una experiencia personal reciente, se elige como problemática a tratar con un sistema multiagente la atención al cliente.
La atención al cliente ha evolucionado con la introducción de modelos de lenguaje avanzados. Los chatbots y asistentes virtuales basados en LLM han demostrado ser eficaces para manejar consultas comunes, pero enfrentan desafíos cuando se trata de interacciones más complejas y específicas del dominio.
La problemática principal es la necesidad de mejorar la velocidad, comprensión y personalización en la atención al cliente en línea. Los usuarios estarán esperando respuestas rápidas y precisas, incluso cuando realicen consultas complejas que requieren un conocimiento más profundo del contexto desde el que la persona pregunta.

- Solución Propuesta:

Se propone un sistema multiagente que aborde los desafíos de la atención al cliente. Cada agente se especializará en una tarea específica, trabajando de manera colaborativa para ofrecer respuestas más rápidas, comprensivas y personalizadas.

- Agentes Involucrados:

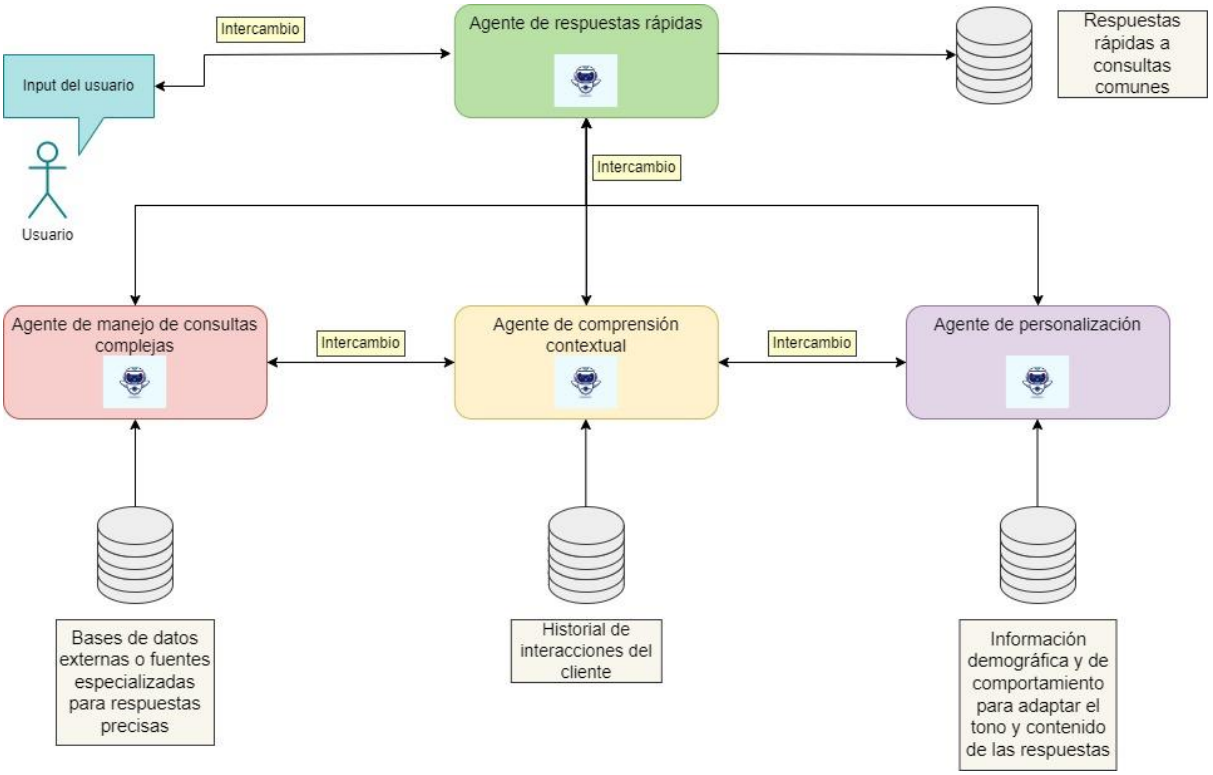
  Agente de respuestas rápidas:
  - Responsabilidad: Proporcionar respuestas rápidas a consultas comunes.Será el agente que devuelva la respuesta al usuario y el primero en ser consultado.
  - Tareas:
    - Identificar patrones en consultas frecuentes y generar respuestas predefinidas.
    - Optimizar la velocidad de respuesta para mejorar la eficiencia del servicio.

  Agente de comprensión contextual:
  - Responsabilidad: Comprender el contexto específico de cada interacción.
  - Tareas:
    - Analizar el historial de interacciones del cliente para contextualizar las consultas.
    - Integrar información sobre preferencias y comportamientos anteriores.

  Agente de manejo de consultas complejas:
  - Responsabilidad: Manejar consultas más complejas que requieren un conocimiento más profundo.
  - Tareas:
    - Identificar consultas que requieren información detallada o conocimientos específicos.
    - Consultar bases de datos externas o fuentes especializadas para respuestas precisas.

  Agente de personalización:
  - Responsabilidad: Personalizar respuestas para adaptarse a las necesidades y preferencias individuales de los clientes.
  - Tareas:
    - Utilizar información demográfica y de comportamiento para adaptar el tono y contenido de las respuestas.
    - Aprender y mejorar continuamente la personalización a lo largo del tiempo.

- Esquema del sistema multiagente:



El enfoque multiagente propuesto intenta considerar y lograr abordar las complejidades de la atención al cliente al aprovechar las fortalezas específicas de cada agente. La colaboración entre agentes especializados permite una atención al usuario más eficaz y personalizada, mejorando la satisfacción y experiencia del mismo.

Conclusiones:

El proceso de generación de un chatbot experto utilizando RAG ha implicado una combinación de preprocesamiento de datos, selección y almacenamiento de información relevante, implementación de modelos para selección de fuentes de contexto y modelos de generación y recuperación.

Junto a ello, se ha realizado posteriormente una evaluación iterativa del desempeño del chatbot mediante análisis de las respuestas brindadas bajo los parámetros definidos, tales como los enviados en la request POST al modelo LLM (temperature, top p, top k), el umbral establecido para determinar la fuente externa a utilizar, el número máximo de resultados solicitados la la base de datos de vectores y la base de datos de grafos.

Las consecutivas consultas realizadas al modelo se han encaminado en el objetivo de hallar el equilibrio entre obtener una respuesta coherente y contextualizada, el tiempo de espera de la respuesta y el uso de recursos de cómputo. Se ha encontrado que es clave iterar y ajustar continuamente el sistema para mejorar su capacidad de respuesta y precisión.

En cuanto a las aplicaciones de agentes inteligentes y sistemas multiagente, la formación de Modelos de Lenguaje Grande (LLM) es un proceso esencial que requiere una atención

meticulosa a los detalles, destacando la importancia de conjuntos de datos diversos y extensos, así como la elección adecuada de la arquitectura del modelo. Estos LLM han catalizado una transformación significativa en diversas aplicaciones, desde la traducción automática hasta la generación de código. No obstante, surgen desafíos éticos, como el sesgo en los datos y la privacidad, subrayando la necesidad de un enfoque reflexivo y proactivo en el desarrollo de IA responsable. Las tendencias futuras se centran en la eficiencia de los modelos, la capacidad de los LLMs para procesar y comprender diferentes tipos de datos de manera simultánea y la personalización, con un énfasis constante en la ética y la confiabilidad, abordando aspectos como la explicabilidad y la mitigación de sesgos para garantizar la implementación responsable de la inteligencia artificial en el futuro.

Fuentes:

- ¿Qué es un Agente Inteligente? Características, tipos, cómo funciona y aplicaciones (ceupe.com)
- El mundo de los Agentes Inteligentes y su utilización en el mundo real | Blog SEAS
- Agentes inteligentes: qué son, aplicaciones y tipos (ccm.net)
- ¿Qué son los modelos de lenguaje de gran tamaño? - Explicación sobre los LLM de IA - AWS (amazon.com)
- ¿Qué es la IA generativa? - Explicación de la inteligencia artificial generativa - AWS (amazon.com)
- ¿Para qué se Utilizan los Grandes Modelos de Lenguaje? | Blog de NVIDIA
- Transformando-la-ensenanza-con-grandes-modelos-de-lenguaje_-una-experiencia-de-utilizacion-de-la-inteligencia-artificial-en-el-aula.pdf (unnoba.edu.ar)
- Decodificando oportunidades y desafíos para agentes de LLM en IA generativa - Unite.AI
- Artificial intelligence in education | UNESCO
- Guidance for generative AI in education and research - UNESCO Biblioteca Digital
- Agente inteligente (inteligencia artificial) - Wikipedia, la enciclopedia libre
- Open LLMs: https://github.com/eugeneyan/open-llms
- Material brindado por la cátedra de Procesamiento del Lenguaje Natural
- Documentación de las librerías utilizadas para el desarrollo del código:
  - https://python.langchain.com/docs/get_started
  - https://huggingface.co/docs/transformers/index
  - https://www.dbpedia.org/
  - https://www.tensorflow.org/api_docs/python/tf