Tecnicatura Universitaria en Inteligencia Artificial

Procesamiento del Lenguaje Natural (IA4.2)

Trabajo Práctico N° 2

Docentes:

- D'Alessandro, Ariel
- Geary, Alan
- Leon Cavallo, Andrea
- Manson, Juan Pablo

Alumno:

- Aguirre, Fabian A-4516/1

Año: 2023

ENUNCIADO

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

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.

DESARROLLO

Ejercicio 1:

A continuación, se presenta la implementación de un chatbot experto utilizando la técnica RAG (Retrieval Augmented Generation). El chatbot se enfoca en temas relacionados con salud y alimentación, y utiliza diversas fuentes de conocimiento, como documentos de texto y datos numéricos en formato tabular, almacenados en una base de datos vectorial y una base de datos de grafos.

Los archivos utilizados son (115 páginas):

alimentacion-basada-en-plantas.pdf
documento-entornos-escolares-saludables.pdf
etica_y_trasplante.pdf
investigacion-sodio.pdf
tasa_vih_jurid.csv

Descripción del código:

1. Descarga de Datos

Se utiliza la biblioteca gdown para descargar archivos desde Google Drive, y se organiza la estructura de carpetas. Se eliminan carpetas temporales después de la descarga. En la descarga se incluye el archivo .env que contiene el token de HuggingFace que se utilizará posteriormente.

```
url = 'https://drive.google.com/drive/folders/1xLQnsyMX8vCWplS3RXqkVcdRdn42GhSA?usp=sharing'

gdown.download_folder(url, quiet=True, output='tp2-nlp')

carpeta_destino = 'documentos'
if not os.path.exists(carpeta_destino):
    os.makedirs(carpeta_destino)

carpeta_origen = 'tp2-nlp/tp2-nlp-documentos'
for filename in os.listdir(carpeta_origen):
    ruta_origen = os.path.join(carpeta_origen, filename)
    ruta_destino = os.path.join(carpeta_destino, filename)
    shutil.move(ruta_origen, ruta_destino)

shutil.rmtree(carpeta_origen)

carpeta_origen = 'tp2-nlp'
carpeta_destino = '/content/'
for filename in os.listdir(carpeta_origen):
    ruta_origen = os.path.join(carpeta_origen, filename)
    ruta_destino = os.path.join(carpeta_destino, filename)
    shutil.move(ruta_origen, ruta_destino)

shutil.rmtree(carpeta_origen)

sys.path.append('/content/')

print("Archivos descargados con éxito.")
```

2. Procesamiento de Texto

Se utiliza la biblioteca fitz para extraer texto de archivos PDF. El texto se divide en fragmentos utilizando la clase RecursiveCharacterTextSplitter de langchain. Se realiza una limpieza básica del texto, convirtiéndolo a minúsculas y eliminando caracteres especiales.

3. Generación de Embeddings y Almacenamiento en ChromaDB

Se emplea el modelo de embeddings universal-sentence-encoder-multilingual de TensorFlow Hub para vectorizar el texto. Los embeddings junto con otros metadatos se almacenan en una base de datos ChromaDB.

```python
#Creación de la base con CrhomaDB
embed = hub.load("https://tfhub.dev/google/universal-sentence-encoder-multilingual/3")

client = chromadb.Client()
collection = client.get_or_create_collection("all-my-documents")

pdf_directory = "documentos"

pdf_files = [f for f in os.listdir(pdf_directory) if f.endswith(".pdf")]

textos = []
ids_textos = []
fuentes = []

def clean_text(text):
    cleaned_text = text.lower()
    cleaned_text = re.sub(r'[^\w\s.,]', '', cleaned_text)
    return cleaned_text

splitter = RecursiveCharacterTextSplitter(chunk_size=80, chunk_overlap=10)
```

```python
for i, pdf_file in enumerate(pdf_files, start=1):
    pdf_path = os.path.join(pdf_directory, pdf_file)

    with fitz.open(pdf_path) as doc:
        text = ""
        for page_num in range(doc.page_count):
            page = doc[page_num]
            text += page.get_text()

    split_texts = splitter.split_text(text)

    clean_texts = [clean_text(sentence) for sentence in split_texts]

    textos.extend(clean_texts)
    ids_textos.extend([f"doc{i}_{j}" for j in range(1, len(clean_texts) + 1)])
    fuentes.extend([f"fuente{i}" for _ in range(len(clean_texts))])

embeddings = embed(textos).numpy().tolist()

collection.add(
    documents=textos,
    metadatas=[{"source": fuente} for fuente in fuentes],
    ids=ids_textos,
    embeddings=embeddings)
```

4. Creación de un Grafo RDF

Se crea un grafo RDF para representar datos numéricos en formato tabular utilizando la biblioteca rdflib. En este caso, se trata de tasas de VIH por jurisdicción. Es tabla contiene poco registros y esta base se crea solo para aplicar los conocimientos adquiridos en la materia ya que no se esta aprovechando ni el rendimiento , flexibilidad y agilidad de estas bases de datos.

```python
#Creación del grafo de tasa_vih
g = Graph()

n = Namespace("http://example.org/place/")
EX = Namespace("http://example.org/terms/")

csv_path = "tasa_vih_jurid.csv"
df = pd.read_csv(csv_path, encoding='utf-8-sig')

def create_rdf_from_data(graph, row, n, EX):
    subject = row["jurisdiccion"]
    obj = row["tasa_vih"]
    relation = "tiene_tasa_vih"

    subject_uri = URIRef(n + f"entity_{subject}")

    graph.add((subject_uri, EX.jurisdiccion, Literal(subject)))
    graph.add((subject_uri, EX.tasa_vih, Literal(obj)))

for _, row in df.iterrows():
    create_rdf_from_data(g, row, n, EX)

g.serialize("tasa-vih-jurid.rdf", format="turtle")
```

5. Chatbot

Se implementa la lógica del chatbot utilizando varias funciones. El chatbot utiliza tanto la base de datos ChromaDB como el grafo RDF para responder preguntas del usuario. Se agrega al contexto los datos recuperados de la base de datos en grafos. No es lo más óptimo pero de otro modo habría que usar técnicas más avanzadas para determinar de donde obtener los datos para responder la pregunta.

- query_chromadb: Realiza consultas en la base de datos ChromaDB.
- zephyr_instruct_template: Genera un template para la conversación.
- generate_answer: Utiliza un modelo de lenguaje (Zephyr) para generar respuestas basadas en el contexto y la pregunta del usuario.
- prepare_prompt: Prepara el contexto para la generación de respuestas.

```python
#CHATBOT

# Cargar el RDF creado anteriormente
graph = Graph()
graph.parse("tasa-vih-jurid.rdf", format="turtle")

def query_chromadb(query_str: str):
    results = collection.query(
        query_embeddings=embed([query_str]).numpy().tolist(),
        n_results=20
    )
    return results


def zephyr_instruct_template(messages):

    template_str = "{% for message in messages %}"
    template_str += "{% if message['role'] == 'user' %}"
    template_str += "{{ message['content'] }}</s>\n"
    template_str += "{% elif message['role'] == 'assistant' %}"
    template_str += "{{ message['content'] }}</s>\n"
    template_str += "{% elif message['role'] == 'system' %}"
    template_str += "{{ message['content'] }}</s>\n"
    template_str += "{% else %}"
    template_str += "{{ message['content'] }}</s>\n"
    template_str += "{% endif %}"
    template_str += "{% endfor %}"

    template = Template(template_str)

    return template.render(messages=messages)
```

```python
def generate_answer(prompt: str, max_new_tokens: int = 768) -> None:
    try:

        api_key = config('HUGGINGFACE_TOKEN')

        api_url = "https://api-inference.huggingface.co/models/HuggingFaceH4/zephyr-7b-beta"

        headers = {"Authorization": f"Bearer {api_key}"}

        data = {
            "inputs": prompt,
            "parameters": {
                "max_new_tokens": max_new_tokens,
                "temperature": 0.7,
                "top_k": 50,
                "top_p": 0.95
            }
        }

        response = requests.post(api_url, headers=headers, json=data)

        respuesta = response.json()[0]["generated_text"][len(prompt):]

        return respuesta

    except Exception as e:
        print(f"An error occurred: {e}")
```

```python
def prepare_prompt(query_str: str):

    TEXT_QA_PROMPT_TMPL = (
        "La información de contexto es la siguiente:\n"
        "---------------------\n"
        "{context_str}\n"
        "---------------------\n"
        "Dada la información de contexto anterior, y sin utilizar conocimiento previo, responde la siguiente pregunta.\n"
        "Pregunta: {query_str}\n"
        "Respuesta: "
    )
    # Se agrega al contexto el contenido extraido de CrhomaDB
    documents = query_chromadb(query_str)

    context_str = ''

    for documents_lista in documents["documents"]:
        for doc in documents_lista:
            context_str += f"{doc}\n"

    # Se agrega al contexto el contenido del grafo
    q = """
        SELECT ?juris ?tasa_vih WHERE {
            ?juris <http://example.org/terms/tasa_vih> ?tasa_vih .
        }
        """

    results = graph.query(q)
```

```
    for r in results:
        juris = r['juris'].split("entity_")[-1]
        context_str += f"{juris} tiene una tasa de VIH de {r['tasa_vih']}\n"

    messages = [
        {
            "role": "system",
            "content": "Eres un asistente útil que siempre responde con respuestas veraces, útiles y basadas en hechos.",
        },
        {"role": "user", "content": TEXT_QA_PROMPT_TMPL.format(context_str=context_str, query_str=query_str)},
    ]

    final_prompt = zephyr_instruct_template(messages)
    return final_prompt

print('**BIENVENIDOS AL CHATBOT DE SALUD Y ALIMENTACIÓN**\n')
print('Las preguntas puede estar relacionadas con los siguiente temas: \n')
print('-Alimentación basada en plantas\n')
print('Entornos escolares saludables\n')
print('Ética y transplantes\n')
print('Sodio en la alimentación\n')
print('Tasa de VIH por provincias\n')
```

6. Interacción con el Usuario

Se proporciona un bucle de interacción donde el usuario puede ingresar preguntas y obtener respuestas del chatbot. El contexto se actualiza en cada iteración.

```
while True:
    user_query = input("Ingrese su consulta (o 'salir' para salir): ")
    if user_query.lower() == 'salir':
        break

    final_prompt = prepare_prompt(user_query)
    print('Respuesta:')
    print(generate_answer(final_prompt))
    print('------------------------------------------------------------')
```

7. Ejemplos de Preguntas

Se proporcionan ejemplos de preguntas que el chatbot puede manejar, relacionadas con tasas de VIH por provincia, procesos de trasplante de órganos, impactos del exceso de sodio en la alimentación, y requerimientos para personas con alimentación vegetariana.

```
#Ejemplos de preguntas
#¿Cuál es la tasa de VIH de Neuquen?
#¿Cómo es el proceso de transplante de organos?
#¿Qué puede producir en el cuerpo el exceso de consumo de sodio en los alimentos?
#¿Cuáles son los requerimientos para personas con alimentación vegetariana?
```

8. Temas Específicos del Chatbot

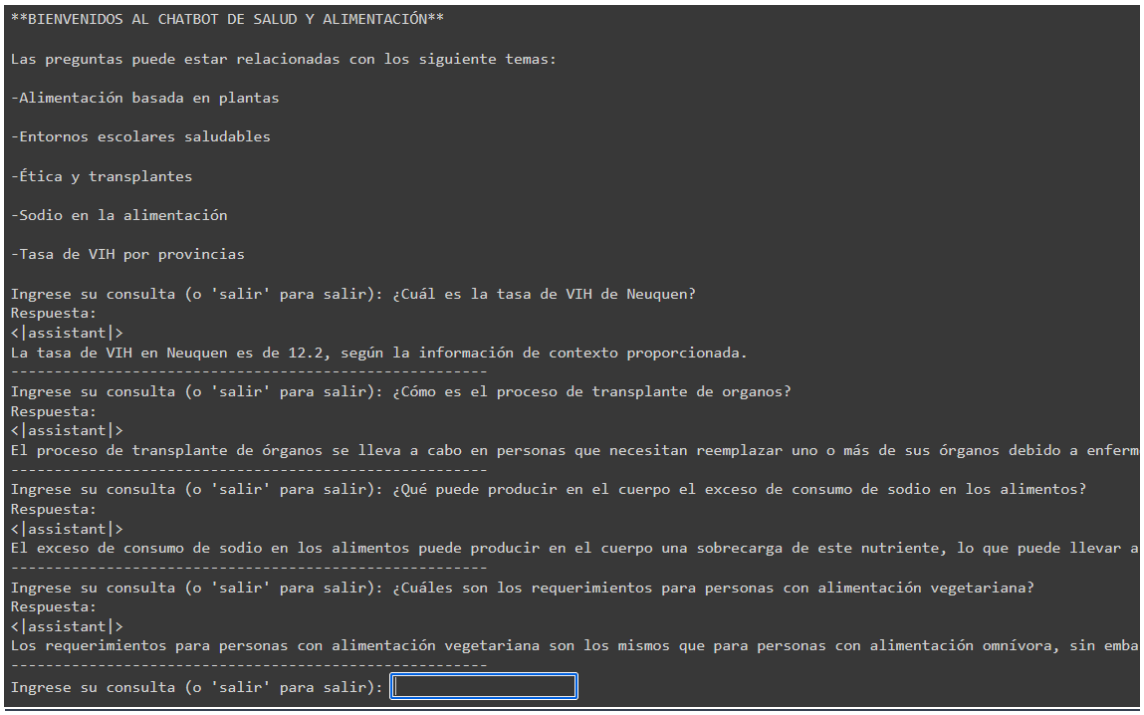
El chatbot se especializa en temas de salud y alimentación, como alimentación basada en plantas, entornos escolares saludables, ética y trasplantes, sodio en la alimentación, y tasas de VIH por provincias.

9. Uso de Modelos Externos

Se utiliza un modelo de lenguaje de Hugging Face (Zephyr) para la generación de respuestas.

10. Finalización del Proyecto

Se muestra un mensaje de bienvenida y se inicia el bucle de interacción con el usuario. El usuario puede ingresar "salir" para terminar la interacción.



Ejercicio 2:

Los Modelos de Lenguaje Grande (LLM) están desempeñando un papel importante en el desarrollo de futuras aplicaciones. Los LLM son muy buenos para comprender el lenguaje debido a la amplia capacitación previa que se ha realizado para los modelos básicos en billones de líneas de texto de dominio público. Métodos como el ajuste fino supervisado y el aprendizaje reforzado con retroalimentación humana (RLHF) hacen que estos LLM sean aún más eficientes para responder preguntas específicas y conversar con los usuarios.

Algunos ejemplos del uso de LLMs son:
Asistentes Virtuales: Aplicaciones como Siri, Google Assistant o Alexa utilizan agentes inteligentes basados en modelos de lenguaje para entender y responder preguntas habladas.

Chatbots en Servicio al Cliente: Empresas utilizan chatbots impulsados por modelos de lenguaje para responder a consultas de clientes en tiempo real.

Generación de Contenido Automático: Herramientas que utilizan modelos de lenguaje para generar automáticamente contenido escrito, como artículos de noticias, resúmenes o descripciones de productos.

Traducción Automática: Sistemas de traducción automática que emplean modelos de lenguaje para traducir texto entre diferentes idiomas.

Análisis de Sentimientos en Redes Sociales: Agentes inteligentes que utilizan modelos de lenguaje para analizar y comprender el sentimiento detrás de publicaciones en redes sociales.

Generación de Texto Creativo: Aplicaciones que generan poesía, historias cortas u otros tipos de contenido creativo mediante modelos de lenguaje.

Corrección Gramatical Automática: Herramientas que utilizan modelos de lenguaje para corregir errores gramaticales y ortográficos en texto.

Asistentes de Escritura: Plataformas que ofrecen sugerencias de escritura y mejoras en el contenido mediante agentes inteligentes basados en modelos de lenguaje.

Generación de Código Automática: Herramientas que asisten en la generación de código fuente a partir de descripciones en lenguaje natural.

Asistentes de Salud Virtual: Agentes inteligentes que proporcionan información de salud, responden preguntas y ofrecen recomendaciones basadas en modelos de lenguaje

Los agentes autónomos, en comparación con simples modelos de chat basados en LLMs (Modelos de Lenguaje con Aprendizaje Profundo), ofrecen mejoras significativas en términos de capacidad para realizar tareas más complejas y adaptarse a entornos dinámicos.

El uso de agentes autónomos se está popularizando en los últimos años, estos son algunos ejemplos:

Drones de Entrega Autónomos: Proyecto Wing (Alphabet, empresa matriz de Google)

Vehículos Autónomos: Waymo (también pertenece a Alphabet)

Tesla Autopilot: Robots de Almacén Autónomos: Kiva Systems (ahora propiedad de Amazon Robotics)

Sistemas de Navegación Autónoma: Sistemas de navegación de robots submarinos para exploración oceánica.

Sistemas de Monitorización Ambiental: Estaciones meteorológicas o de monitoreo de calidad del aire autónomas.

Sistemas de Seguridad Autónomos: Sistemas de vigilancia autónoma basados en IA.

Robots Agrícolas Autónomos: Robótica agrícola para tareas como la cosecha automatizada.

Robots de Exploración Espacial: Rover Curiosity de la NASA en Marte.

Algunos de los proyectos que utilizan LLMs son:

AutoGPT: AutoGPT es un agente autónomo que desarrolla un LLM subyacente para comprender el objetivo que se le ha asignado y trabajar para lograrlo. AutoGPT genera una lista de tareas que cree que necesita para cumplir con lo que le pediste, sin requerir más información o mensajes.

GPT-Engineer: Al igual que AutoGPT, GPT-Engineer es otro agente autónomo que utiliza LLM para comprender y trabajar hacia el objetivo asignado

BabyAGI: Este es otro proyecto que se centra en el desarrollo de agentes autónomos con habilidades de planificación a largo plazo y uso de memoria. BabyAGI busca crear agentes que puedan aprender y adaptarse a su entorno utilizando técnicas de aprendizaje por refuerzo.

También hay proyectos como CAMEL y Generative Agents se centran en la creación de entornos de simulación específicos para agentes autónomos.

Sistema Multiagente para Optimizar el Tráfico Urbano

Problemática: El tráfico urbano ineficiente afecta la calidad de vida de los habitantes, generando congestiones, contaminación ambiental y aumentando el tiempo de viaje. Además, vehículos de prioridad como ambulancias, bomberos o vehículos policiales se ven demorados por el caos del tránsito. La problemática central radica en la falta de coordinación entre los diferentes elementos del tráfico, como semáforos, señales y sistemas de transporte público.

Objetivo: Desarrollar un sistema multiagente que permita la gestión coordinada y eficiente del tráfico urbano, reduciendo la congestión, mejorando los tiempos de viaje y reduciendo los tiempos de vehículos con prioridad.

Desarrollo del Sistema Multiagente:

- **Agente de Central Coordinador:**
  - Coordina la operación de todo el sistema multi agente.
  - Se comunica con todos los agentes incluyendo la información obtenida por el agente usuario.

  Este agente se encarga de coordinar a todos los otros agentes para una respuesta óptima. Esta en constante comunicación con todo el sistema por lo que le permite tener una visión global.

- **Agente Usuario:**
  - Establece la comunicación con el usuario administrador del sistema

Este agente establece una comunicación en lenguaje natural con el administrador para mostrar reportes, indicadores del estado general del tránsito. Además permite la intervención del usuario en todo momento de cualquier parte del sistema

- **Agente de Control de Semáforos:**
  - Utiliza datos en tiempo real para ajustar la duración de los semáforos.
  - Coordinación con los distintos semáforos para establecer "onda verde"
  - Ajusta los tiempos de los semáforos para favorecer los vehículos con prioridad

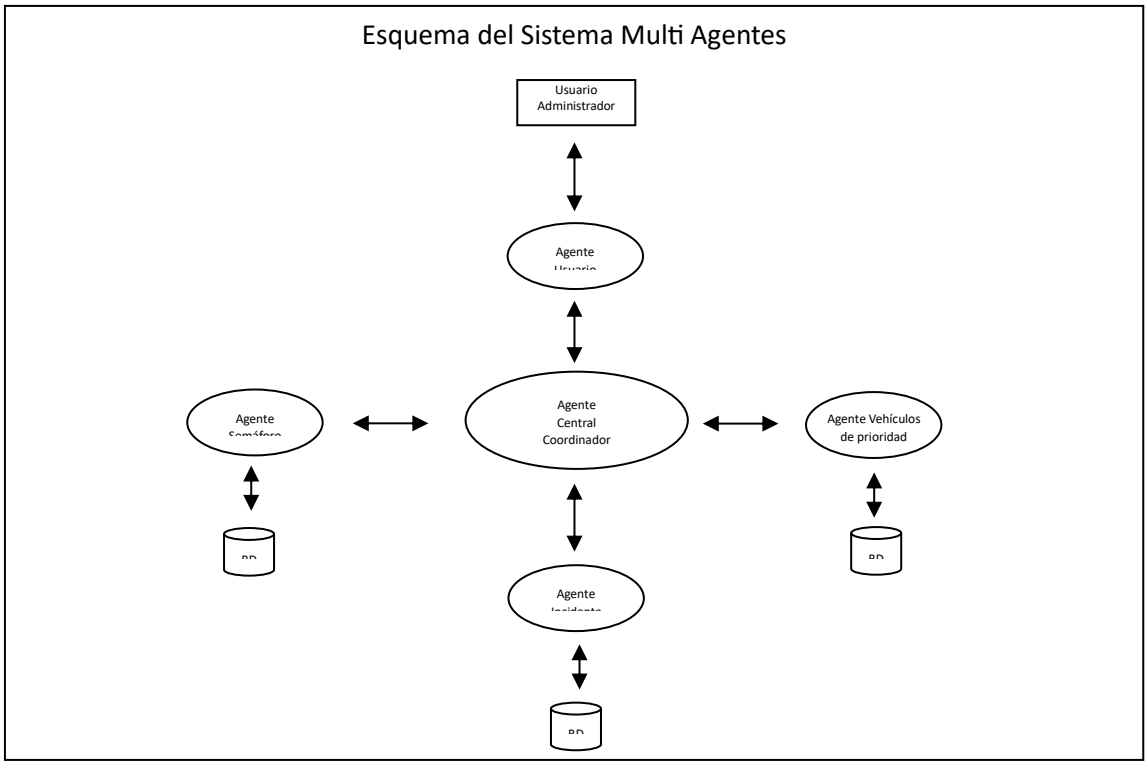
  Este agente se nutre de los datos almacenado en una base de datos (recopilados a través de múltiples sensores) de todos los datos referidos a los semáforos. Toma decisiones referidas a la operatoria todos los semáforos en general.

- **Agente de vehículos de prioridad:**
  - Optimiza las rutas para vehículos de prioridad

  Este agente se nutre de los datos almacenado en una base de datos (recopilados a través de múltiples sensores y datos de otras fuentes) de todos los datos referidos a los vehículos de prioridad. Toma decisiones referidas al aviso de los conductores en general y a los de los vehículos de prioridad.

- **Agente de Respuesta a Incidentes:**
  - Detecta eventos como accidentes, obras viales u otros imprevistos y ajusta la circulación.

  Este agente se nutre de los datos almacenado en una base de datos (recopilados a través de múltiples sensores y cámaras) de todos los datos referidos a cualquier accidente o imprevistos. Toma decisiones referidas al aviso de los conductores y los otros agentes para ajustar la circulación.



Implementación de Tecnologías: El sistema utilizará tecnologías emergentes como el Internet de las Cosas (IoT), análisis de big data y aprendizaje automático para recopilar y analizar datos en tiempo real.

Beneficios Esperados:

- Reducción de tiempos de viaje para los ciudadanos.
- Minimización de emisiones de gases contaminantes.
- Mejora en la eficiencia de los vehículos de prioridad.

Conclusiones: La implementación de un sistema multiagente para la gestión del tráfico en una ciudad inteligente es crucial para abordar la problemática del tráfico urbano. La coordinación entre diferentes agentes permite una respuesta dinámica a situaciones cambiantes, mejorando la eficiencia y la sostenibilidad del sistema de transporte urbano. La importancia de administrar de una forma óptima las rutas de los vehículos mejora enormemente la fluidez del tránsito

Fuentes:

https://en.wikipedia.org/wiki/Wing_%28company%29

https://es.wikipedia.org/wiki/Waymo

https://www.xataka.com/automovil/waymo-sera-la-nueva-compania-independiente-de-alphabet-google-encargada-de-coches-autonomos

https://www.tesla.com/autopilot

https://en.wikipedia.org/wiki/Amazon_Robotics

https://invdes.com.mx/tecnologia/los-robots-marinos-que-exploran-aguas-profundas-del-golfo-de-mexico/

https://www.smn.gob.ar/noticias/nuevas-estaciones-autom%C3%A1ticas

https://www.soyseguridadprivada.com/inteligencia-artificial-en-la-seguridad-privada-de-los-proximos-10-anos/

https://www.edsrobotics.com/blog/agricultura-automatizada-y-robotica-agricola/

https://spaceplace.nasa.gov/mars-curiosity/sp/

https://autogpt.net/

https://babyagi.org/docs/README-es.html

https://github.com/AntonOsika/gpt-engineer

https://medium.com/latinxinai/agentes-aut%C3%B3nomos-y-simulaciones-en-llm-un-vistazo-a-autogpt-babyagi-camel-y-generative-agents-bdb0bbfcddac

Unidad 6 - Chatbots y Sistemas de Diálogo (TUIA-NLP)

Unidad 7 - Agentes Autónomos y Sistemas Inteligentes (TUIA-NLP)