Aprovechamiento de Modelos de Lenguaje para la Visualización Automática de Datos

Macarena Repetto[1,*] y Marcos Zárate[1,2]

[1] Laboratorio de Investigaciones en Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco (LINVI-UNPSJB), Puerto Madryn, Argentina.
macarepetto@gmail.com

[2] Centro para el Estudio de Sistemas Marinos, Centro de Investigación Nacional Patagónico (CESIMAR-CENPAT-CONICET), Puerto Madryn, Argentina.
zarate@cenpat-conicet.gob.ar

Resumen Este breve artículo explora las capacidades de LIDA, una herramienta para la generación automática de visualizaciones e infografías de forma agnóstica a la gramática, utilizando Modelos de Lenguaje Grande (LLMs). Al abordar las limitaciones de las herramientas de visualización tradicionales, LIDA proporciona una solución flexible, agnóstica a la gramática, que integra el procesamiento del lenguaje natural con la generación de visualizaciones. A través de un caso de estudio relacionado con mamíferos marinos generado dentro del CESIMAR-CONICET, demostramos el proceso de múltiples etapas de LIDA para resumir, explorar objetivos, generar visualizaciones y crear infografías.

Palabras clave: Visualización de Datos, Data Visualization, LLMs, LIDA

1. Introducción y motivación

La visualización de datos [1] es un aspecto crucial del análisis de datos, al permitir la transformación de conjuntos de datos complejos en formatos visuales comprensibles, facilita el descubrimiento de patrones, tendencias y conocimientos que podrían permanecer ocultos en los datos crudos. Con el advenimiento de los LLMs [2], existe una oportunidad significativa para automatizar y mejorar el proceso de visualización, haciéndolo más accesible y eficiente. Los LLMs, como GPT-4[3], han demostrado capacidades notables en la comprensión y generación de lenguaje natural abriendo nuevas posibilidades para su aplicación en la visualización de datos científicos.

Sin embargo, crear visualizaciones efectivas requiere más que solo generar gráficos y diagramas. Implica entender la semántica de los datos, identificar objetivos de visualización relevantes y generar especificaciones de visualización

* Corresponding author.

[3] https://openai.com/index/gpt-4/

adecuadas. Este proceso puede ser desafiante y consumir mucho tiempo, particularmente para los usuarios que pueden no tener una amplia experiencia en técnicas de visualización de datos. Las herramientas de visualización tradicionales a menudo dependen de gramáticas predefinidas y requieren que los usuarios tengan un profundo conocimiento de sus respectivas sintaxis y bibliotecas. Esta curva de aprendizaje puede ser una barrera significativa, especialmente para aquellos que son nuevos en la ciencia de datos o que necesitan generar rápidamente conocimientos a partir de grandes conjuntos de datos.

LIDA [3] aborda estos desafíos aprovechando los LLMs y los modelos de generación de imágenes (IGMs) para automatizar la generación de visualizaciones agnósticas de la gramática. Los métodos tradicionales a menudo requieren un esfuerzo manual significativo y experiencia, haciendo que el proceso de visualización lleve mucho tiempo y sea propenso a errores humanos. La integración de los LLMs con herramientas de visualización ofrece un camino hacia flujos de trabajo de análisis de datos más intuitivos y eficientes. Al entender consultas en lenguaje natural, los LLMs pueden interpretar las necesidades del usuario y traducirlas en comandos de visualización precisos. Este enfoque no solo simplifica el proceso para usuarios con diferentes niveles de experiencia, sino que también mejora la calidad y relevancia de las visualizaciones generadas.

La arquitectura de LIDA está diseñada para facilitar una interacción fluida entre el usuario y la herramienta de visualización. Comprende de cuatro módulos principales: 1) `Summarizer`, 2) `Goal Explorer`, 3) `VisGenerator` y 4) `Infographer`. Summarizer convierte datos crudos en resúmenes concisos en lenguaje natural, que sirven como base para las tareas de visualización subsecuentes. El Goal Explorer identifica posibles objetivos de visualización basados en estos resúmenes, mientras que VisGenerator crea y refina el código de visualización en varios lenguajes de programación y gramáticas. Finalmente, Infographer genera infografías fieles a los datos utilizando IGMs, asegurando que las visualizaciones no solo sean precisas, sino también visualmente atractivas. Al automatizar estos pasos, LIDA reduce el esfuerzo manual requerido para crear visualizaciones efectivas y permite a los usuarios centrarse en interpretar los resultados y tomar decisiones basadas en datos. Este artículo breve explora la arquitectura y funcionalidad de LIDA, destacando su potencial para transformar la forma en que se crean y utilizan las visualizaciones de datos en la ciencia marina.

2. Estado del arte

Las herramientas de visualización de datos actuales a menudo dependen de gramáticas predefinidas y requieren una intervención manual extensa. Herramientas tradicionales como Matplotlib [4], Seaborn [5] y D3.js [6] son poderosas pero requieren un profundo conocimiento de su sintaxis y bibliotecas respectivas. Los avances recientes han introducido interfaces más intuitivas y funciones de automatización, pero aún no manejan eficientemente gramáticas de visualización diversas y datos a gran escala. LIDA se destaca al ofrecer una solución integral que integra LLMs para el procesamiento del lenguaje natural y IGMs para crear

infografías estilizadas. Este enfoque de múltiples etapas asegura que LIDA pueda manejar una amplia gama de tareas de visualización, desde gráficos simples hasta infografías complejas y personalizadas. Al ser agnóstico de la gramática, LIDA admite múltiples marcos de visualización, incluidos Altair, Matplotlib e incluso bibliotecas que no sean de Python, lo que lo convierte en una herramienta versátil para científicos y analistas de datos. Además, la capacidad de LIDA para interpretar y generar visualizaciones en respuesta a consultas en lenguaje natural lo distingue de otras herramientas.

3. Caso de Estudio: Distribución de mamíferos marinos

Adquisición de Datos: Para llevar a cabo nuestros experimentos, hemos utilizado el conjunto de datos titulado "Distribución de elefantes marinos del sur en Península Valdés", disponible públicamente a través del DOI: 10.15468/gx7bcw. El conjunto de datos utilizado en este caso de estudio proviene de la Global Biodiversity Information Facility (GBIF) y se centra en los elefantes marinos, abarcando desde 2002 hasta 2022. Contiene un total de 5432 registros y se pueden diferenciar las siguientes columnas:

- **Fecha:** La fecha en que se registraron los datos.
- **Cuaderno:** Número de libreta de campo utilizado para anotar los datos.
- **Orden:** El orden en que se registraron los datos.
- **Cantidad de Hembras:** El número de elefantes marinos hembra observados.
- **Cantidad de Machos:** El número de elefantes marinos macho observados.
- **Hembras por Grupo:** El número promedio de hembras por grupo.
- **Referencia:** Indica si el harén tiene o no un macho alfa.
- **Playa:** El nombre de la playa donde se realizó la observación.
- **Longitud:** La coordenada longitudinal del lugar de observación.
- **Latitud:** La coordenada latitudinal del lugar de observación.
- **Tipo:** El tipo de censo que se realizo.

Este conjunto de datos proporciona información valiosa sobre la dinámica poblacional, distribución y comportamiento de los elefantes marinos a lo largo de un período de 20 años. Los datos recogidos incluyen información detallada sobre tanto las ubicaciones geográficas como las características demográficas de las poblaciones de elefantes marinos observadas, convirtiéndolo en un recurso valioso para estudios ecológicos y ambientales.

Conectar LIDA a OpenAI, Resumir y Generar Objetivos: Para iniciar el proceso, se conectó LIDA a la API de OpenAI utilizando una clave API disponible. El siguiente paso fue resumir los datos, LIDA utilizó el conjunto de datos y generó un resumen en lenguaje natural que describía la información clave de cada columna. (Ver Fig. 1)

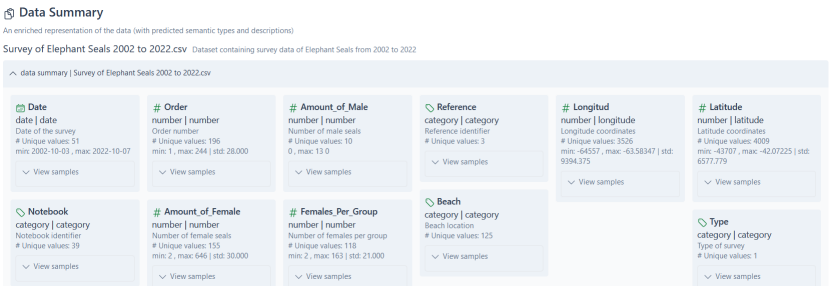

Figura 1. Resumen de Datos Generado por LIDA

A continuación, se generaron los objetivos de visualización, en este caso LIDA identificó varios objetivos potenciales basados en el resumen: 1) Distribución de elefantes marinos hembra a lo largo de los años, 2) Relación machos y hembras de elefantes marinos, 3) Distribución de ejemplares en las diferentes playas, 4) Tendencia de la cantidad de hembras por harem a través de los años, 5) Distribución del orden del relevamiento.

Generar Gráfico en base a un objetivo: Con los objetivos establecidos, el modulo VisGenerator de LIDA generó automáticamente el código de visualización en Seaborn o cualquiera de las herramientas seleccionadas. Uno de los objetivos era mostrar la distribución de elefantes marinos hembra a lo largo del tiempo y otro mostrar la relación entre la cantidad de machos y hembras. (Ver Fig. 2)

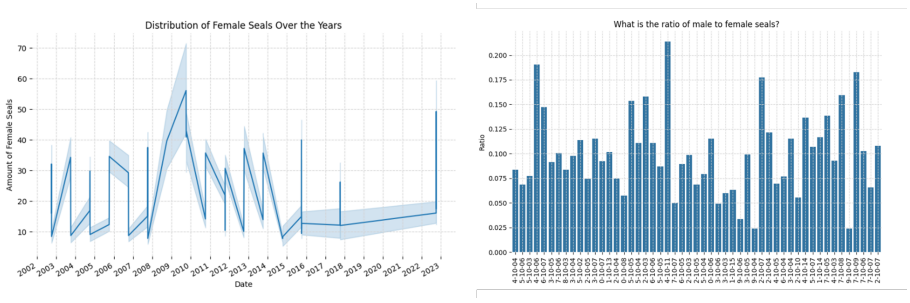

Figura 2. A la izquierda se puede ver la distribución de hembras (Objetivo 1) sugerido por LIDA, mientras que a la derecha se puede ver la relación machos/hembras (Objetivo 2).

Generar y Redefinir una visualización mediante una consulta del usuario: Usando lenguaje natural, ingresamos el siguiente comando para instruir a LIDA: "Mostrame la distribución de elefantes marinos machos a lo largo de los años" y como se muestra en la Fig. 3 LIDA realizó el gráfico correspondiente. Además, el usuario puede solicitar mejoras en las visualizaciones generadas a través de consultas adicionales en lenguaje natural, como por ejemplo: cambiar el título, los colores o las etiquetas.

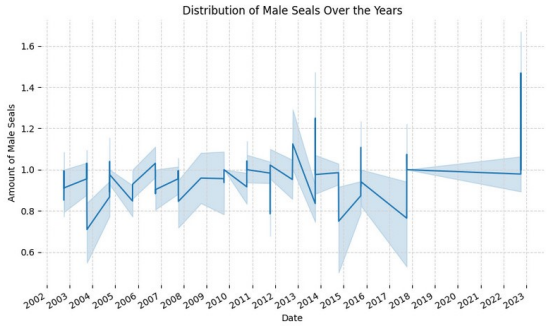

Figura 3. Distribución de machos como fue requerido por el usuario.

Todas las pruebas realizadas en este trabajo fueron llevadas a cabo clonando el repositorio oficial de LIDA y utilizando una API KEY paga de GPT4.

4. Conclusión y Trabajo a Futuro

LIDA representa un avance significativo en el campo de la visualización automatizada de datos. Al aprovechar las capacidades de los modelos LLMs e IGMs, aborda las limitaciones de las herramientas tradicionales y proporciona una solución flexible y agnóstica en cuanto a gramática para generar visualizaciones de alta calidad. La integración del procesamiento de lenguaje natural con la generación de visualizaciones permite a los usuarios interactuar con los datos de manera más intuitiva y eficiente, reduciendo la necesidad de codificación manual extensa y conocimientos especializados. El trabajo futuro se centrará en expandir el soporte de bibliotecas de LIDA, mejorar su manejo en conjuntos de datos provenientes de las ciencias marinas e integrar características más avanzadas para la visualización interactiva y en tiempo real de datos oceanográficos.

Además, la incorporación de retroalimentación de los usuarios y técnicas de aprendizaje automático para mejorar continuamente la calidad y relevancia de las visualizaciones generadas aumentaría aún más la utilidad de LIDA.

Referencias

1. Frits H Post, Gregory Nielson, and Georges-Pierre Bonneau. Data visualization: The state of the art. 2002.
2. Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
3. Victor Dibia. Lida: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. March 2023.
4. J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.
5. Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.
6. Mike Bostock. D3.js - data-driven documents, 2012.