UNIVERSIDAD DE GRANADA

FACULTAD DE CIENCIAS

Propuesta TFG. Curso 2024/2025

GRADO: Grado en Estadística

CÓDIGO DEL TFG: 223-026-2024/2025

1. DATOS BÁSICOS DEL TFG:

Título: Bases estadísticas de modelos de lenguaje de gran tamaño (LLM) como ChatGPT.

Descripción general (resumen y metodología):

En los últimos años, modelos de procesamiento del lenguaje natural como GPT, Gemini o Claude, están siendo utilizados por el público general con gran popularidad. Generalmente, estos modelos se suelen entender como una caja negra donde la "inteligencia" que contiene es capaz de realizar tareas que hasta hace muy poco tiempo eran impensables para una máquina. Sin embargo, las bases teóricas que fundamentan estos modelos son principalmente estadísticas.

Según el propio ChatGPT: "ChatGPT, desarrollado por OpenAI, se fundamenta en técnicas avanzadas de estadística y aprendizaje automático, específicamente mediante redes neuronales profundas. Utiliza modelos transformadores, como el GPT-4, que emplean algoritmos estadísticos para procesar y aprender patrones a partir de extensos corpora textuales. Durante el entrenamiento, se optimizan miles de millones de parámetros utilizando métodos de inferencia estadística y ajuste de modelos, lo cual permite al sistema calcular la probabilidad condicional de secuencias de palabras y generar respuestas coherentes y contextualmente adecuadas. Así, ChatGPT aplica principios estadísticos avanzados para interpretar y producir lenguaje natural de manera precisa y efectiva."

En este proyecto se pretende profundizar en las bases estadísticas de estos modelos, así como analizar la potencialidad y limitaciones que intrínsecamente conlleva el uso de estas técnicas.

Tipología: Trabajos bibliográficos sobre el estado actual de una temática relacionada con el Grado.

Objetivos planteados:

- Realización de un Análisis Bibliográfico sobre la estadística en los LLM.
- Identificación de la potencialidad y las limitaciones de los LLM.
- Familiarización con el análisis de datos no estructurados, como es el procesamiento del lenguaje natural.

Bibliografía básica:

- Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. **"Attention is All you Need."** Neural Information Processing Systems. 2017.
- Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. **"Improving language understanding by generative pre-training."** 2018.
- Zong, C., Xia, R., & Zhang, J. **"Text data mining."** Singapore: Springer. 2021.
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. **"An Introduction to Statistical Learning with Applications in R."** Springer, 2nd ed. 2021.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman. **"The Elements of Statistical Learning: Data Mining, Inference, and Prediction."** Springer; 2nd ed. 2009.

Recomendaciones y orientaciones para el estudiante:

Se recomienda haber cursado las asignaturas "Minería de Datos" y "Técnicas Avanzadas de Estadística Multivariante".

Plazas: 1

2. DATOS DEL TUTOR/A:

Nombre y apellidos: FRANCISCO JAVIER ARNEDO FERNÁNDEZ

Ámbito de conocimiento/Departamento: ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Correo electrónico: arnedo@ugr.es

3. COTUTOR/A DE LA UGR (en su caso):

Nombre y apellidos:

Ámbito de conocimiento/Departamento:

Correo electrónico:

4. COTUTOR/A EXTERNO/A (en su caso):

Nombre y apellidos:

Correo electrónico:

Nombre de la empresa o institución:

Dirección postal:

Puesto del tutor en la empresa o institución:

5. DATOS DEL ESTUDIANTE:

Nombre y apellidos:

Correo electrónico: