Uso de modelos de lenguaje de gran tamaño (LLM) de código abierto

¿Es posible hacerlo "On Premise"?

Isaac Vidal Daza

Apoyo a la Docencia Centro de Servicios de Informática y Redes de Comunicaciones Universidad de Granada

28-05-2024





Teaching Support Department

Members

- Francisco Romera Juárez. (Head)
- Fernando López Álvarez.
- José Guerrero Peregrina.
- Antonio Cano Ruano.

- Rodrigo González Gálvez.
- Domingo Baca Ruíz.
- Leire Melchor López.
- Isaac Vidal Daza.



Teaching Support Department

Services

- Computers Classrooms Management.
- Software Deployment.
- Virtual Desktop Infrastructure.
- Microsoft 365 Management.
- Systems and Services Deployment.

Inventory

- 25 Faculties (Ceuta y Melilla).
- 128 Computers Classrooms.
- 3418 PC's.
- 200 Virtual Desktops.

Users

- 53000 Students.
- 3600 Teaching Staff.
- 2200 Administration Staff.

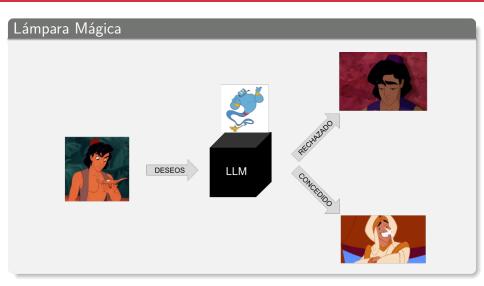


Inteligencia Artificial (según los "Mass Media")





Modelos de Lenguaje





Modelos de Lenguaje





¿Cómo pasar de Espectador a Ilusionista?

Utilizando modelos LLM Open Source

- Llama3*
- Mistral

¿Dónde se encuentran?

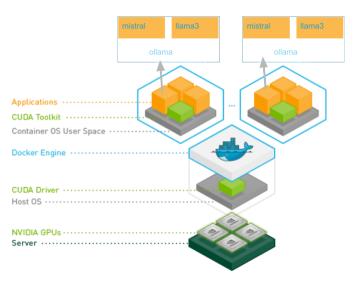
- https://huggingface.co/
- https://mistral.ai/

¿Dónde ejecutan?

- On Premise
- Nube



¿Cómo ejecutarlos? CPUs y GPUs





Inconvenientes y Aprendizaje

¿No hay que hacer inversión millonaria?





Interfaz Gráfica (unas 40 líneas de código)

Listing 1: Código usando el framework Chainlit

```
@cl.on_chat_start
1
          async def on chat start():
             model = Ollama(base_url="http://localhost:11434", model="llama3")
             prompt = ChatPromptTemplate.from_messages(
                         "system",
                            "You are a helpful assistant. You must always make responses to
                                   spanish
                            question, if you recieve a question in any other language.
9
                                  Moreover, you
                            have to answer in spanish."
10
                     ("human", "{question}"),
13
14
              runnable = prompt | model | StrOutputParser()
15
              cl.user session.set("runnable", runnable)
16
```



Interfaz Gráfica (unas 40 líneas de código)

Listing 2: Código usando el framework Chainlit



Resultado 40 líneas de código

¿No hay que hacer inversión millonaria?

