

Bienvenidos



Objetivo

Crear Comunidad de Aprendizaje



Agenda

- Bienvenida de Andrew NG
- En nuestros equipos:
 - Ejecutar LLMs
 - Invocar c/LLMs desde un notebook Python
 - Hacer búsqueda vectorial
 - Ejecutar nuestro primer RAG

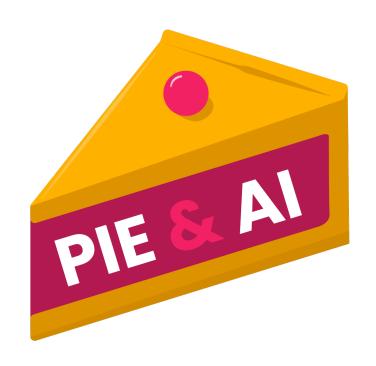
 Experiencia práctica y cuellos de botella en iniciativas IA Generativa

en la vida real



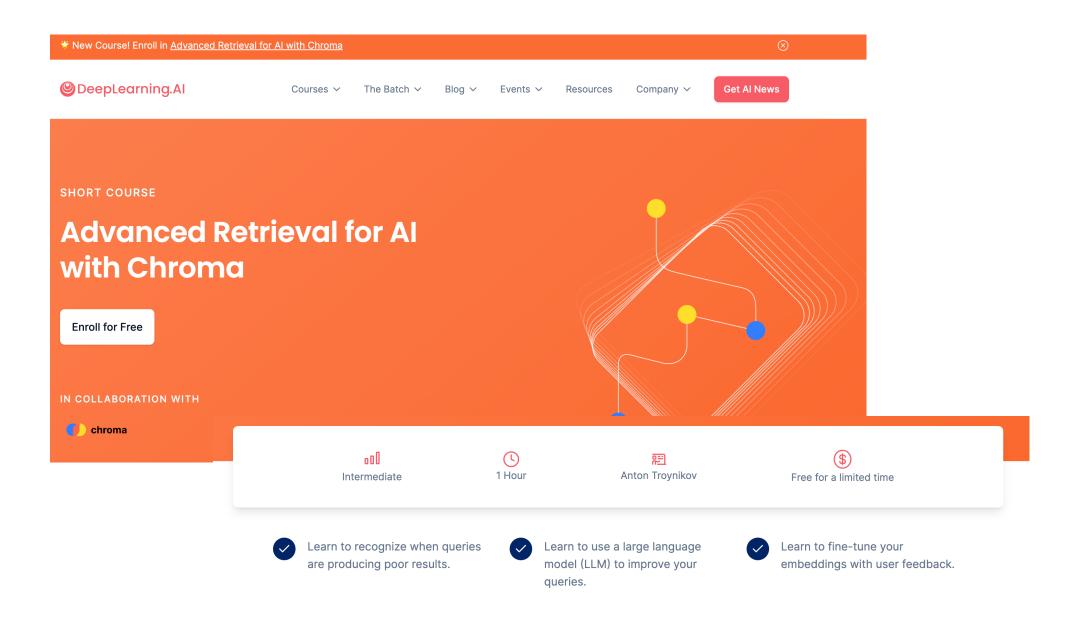






Building RAG Applications





Agenda

- En nuestros equipos:
 - Ejecutar LLMs
 - Invocar c/LLMs desde un notebook Python
 - Hacer búsqueda vectorial
 - Ejecutar nuestro primer RAG

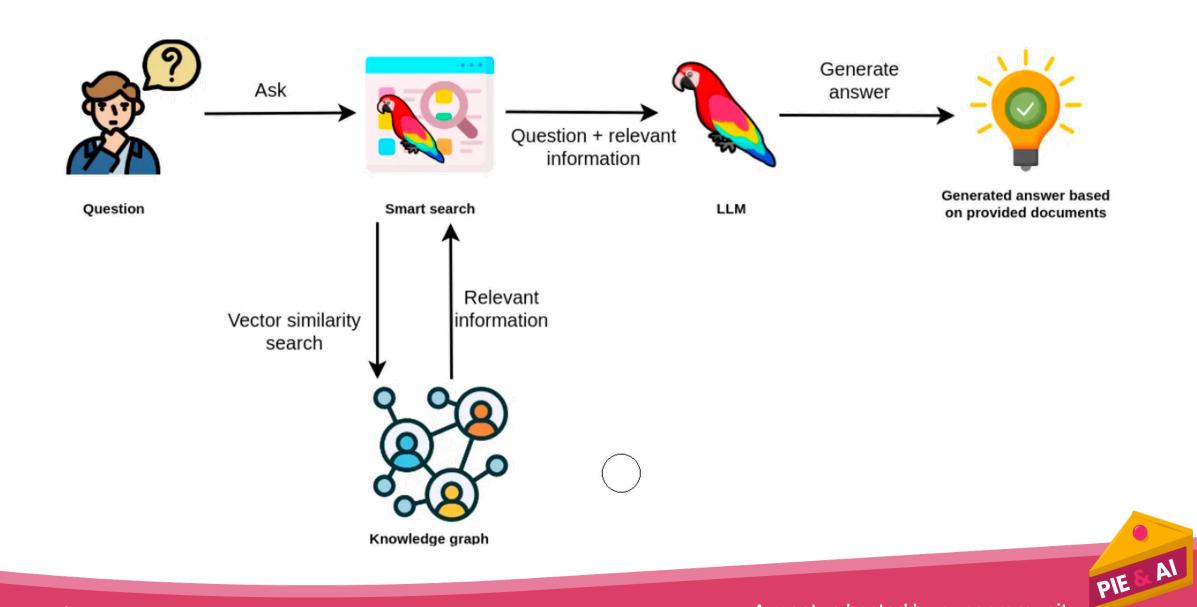
• Experiencia práctica y cuellos de

botella en iniciativas IA Generativa

en la vida real



https://blog.langchain.dev/using-a-knowledge-graph-to-implement-a-devops-rag-application/



LLM

DeepLearning.Al



10 TB Texto

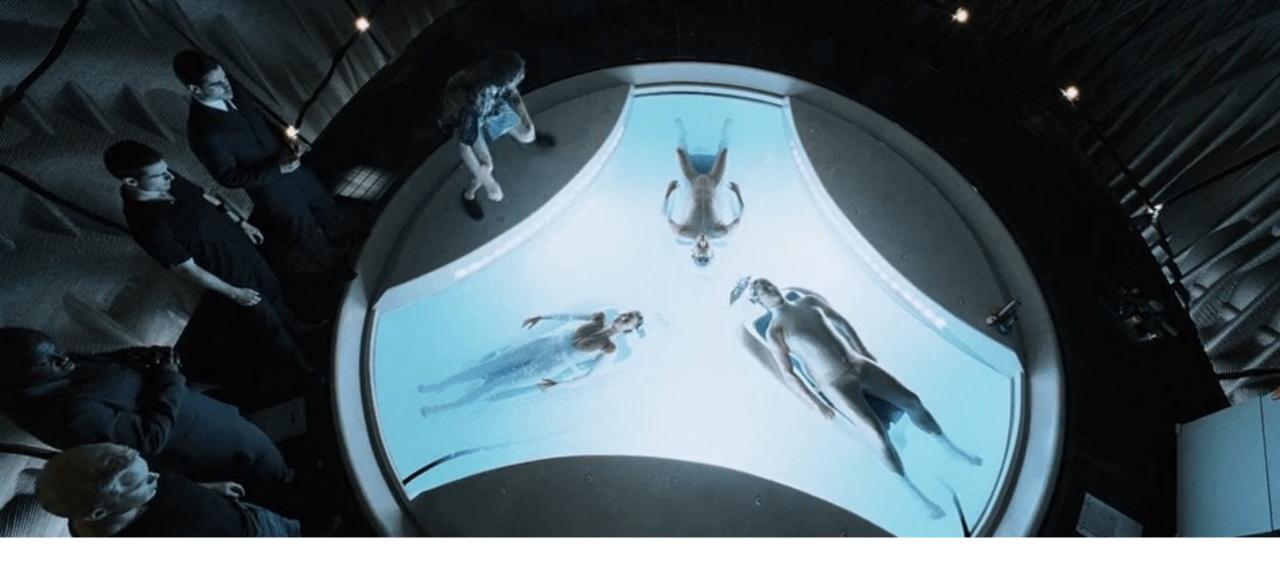


6000 GPUs x 12 días (2M USD)

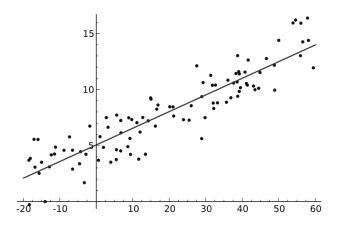


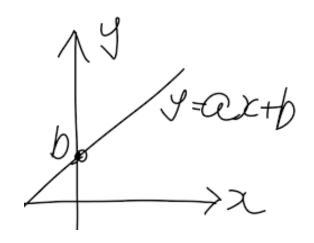
140 GB

Intro to Large Language Models, Andrej Karpathy

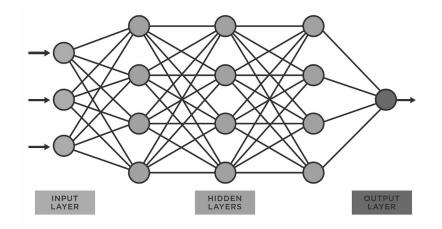










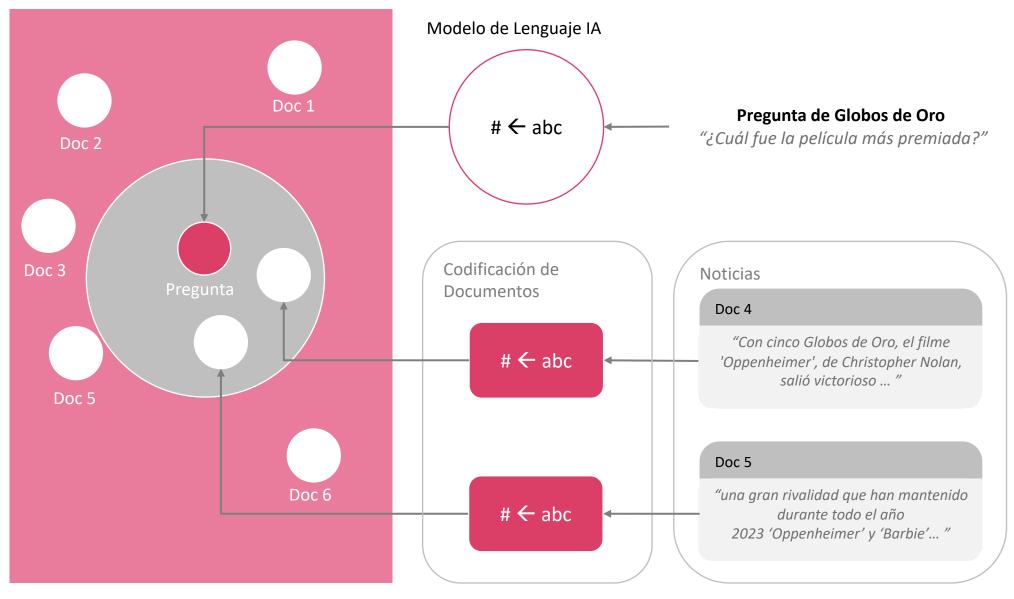


Trained to predict the next word in a sentence:

The cat is chasing the

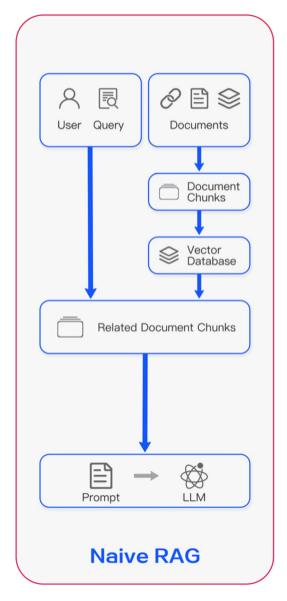
dog 5%
mouse 70%
squirrel 20%
boy 5%
house 0%

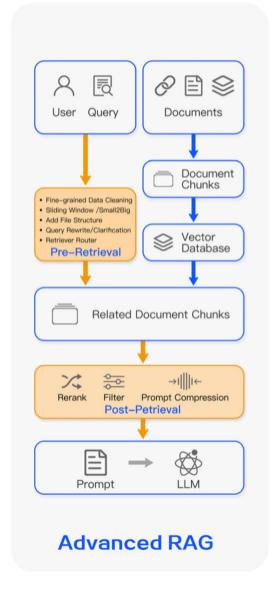


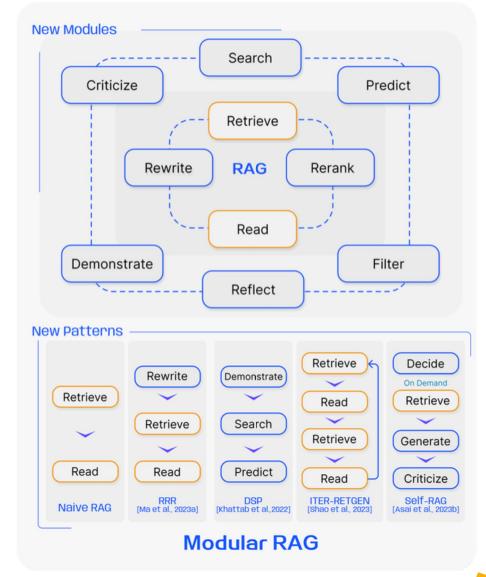


Adaptado de "Retrieval Augmented Generation with Huggingface Transformers and Ray" https://huggingface.co/blog/ray-rag









Retrieval-Augmented Generation for Large Language Models: A Survey

YunfanGao1, YunXiong2, XinyuGao2, KangxiangJia2, JinliuPan2, YuxiBi3, Yi Dai1, Jiawei Sun1 and Haofen Wang 1,3 *



Agenda

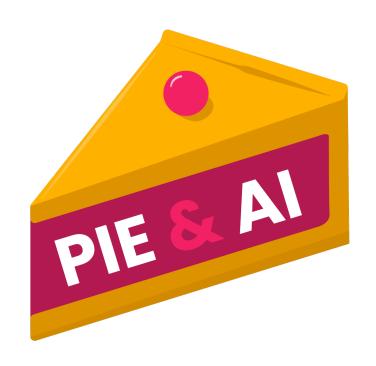
- En nuestros equipos:
 - Ejecutar LLMs
 - Invocar c/LLMs desde un notebook Python
 - Hacer búsqueda vectorial
 - Ejecutar nuestro primer RAG

• Experiencia práctica y cuellos de

botella en iniciativas IA Generativa

en la vida real





Building RAG Applications

