RAG (Retrieval-Augmented Generation) - BASE DE DATOS VECTORIAL

1. Creación de BBDD vectorial:

* A partir de con título, abstract, categoría y id de los primeros 50.000 registros del dataset original
* Se crean embeddings de título + abstract con modelo sentence-transformers (all-MiniLM-L6-v2)
* Se crea un nuevo índice FAISS del tipo "Flat" (Crea un nuevo índice FAISS del tipo "Flat" usando la distancia L2 (distancia euclidiana).
  + FAISS ofrece varios tipos de índices. FlatL2 y FlatIP son los más precisos (búsquedas exactas). El segundo utiliza la similitud coseno en lugar de la distancia euclídea.
* Se normalizan los embeddings y se añaden al índice de búsqueda FAISS[1](https://medium.com/loopio-tech/how-to-use-faiss-to-build-your-first-similarity-search-bf0f708aa772)
* Se crean 2 ficheros: arxiv\_data.pkl (con con título, abstract, categoría y id de los primeros 50.000 registros en formato pickle) y arxiv\_index.faiss (con los embeddings normalizados de los resúmenes en un formato optimizado de FAISS)

1. Búsqueda en la BBDD vectorial:

* La consulta se convierte en un embedding (con mismo modelo sentence-transformer) y se normaliza (de la misma forma que los embeddings de la BBDD lo están)
* Se usa el índice FAISS para encontrar los vectores más cercanos (distancia euclídea) de entre el total de documentos en la BBDD
* Se transforma la distancia en un valor de similitud utilizando la transformación exponencial: similitud = e^-alpha\*dist. Las ventajas de usar la transformación exponencial son:
  + Tiene fundamento en teoría de probabilidad y estadística
  + El parámetro alpha te da control sobre la "severidad" de la transformación
  + Los valores siempre estarán entre 0 y 1
* Se establece un umbral de similitud de 0.75 a partir del cual se considera que 2 vectores son similares
* Se usa el archivo PKL para recuperar el título, abstract, etc. asociados a esos vectores

Nota: utilizamos faiss-cpu porque faiss-gpu no está disponible directamente a través de pip en Windows. Sin embargo, la versión CPU es más que suficiente para este caso de uso en que trabajamos con 50.000 vectores.

Hacer diagrama como este para explicarlo en la Presentación:  
Diagrama

Descripción generada automáticamente