Universidad de Oviedo

Prácticas de laboratorio

Sesión 6

Sistemas de Información para la Web

Índices densos

Omar Teixeira González, *UO281847*

23/11/2023

Tabla de contenidos

Documentación sobre Índices Densos

Explicación del código realizado

Para empezar, se han creado los índices de la forma vista en la práctica 4 y en la clase de laboratorio, sin embargo, se han encontrado un par de errores con respecto a la parte de creación del índice denso.

El primer error en cuestión viene relacionado con el mencionado en el documento, correspondiente al uso de *use_ann*, o búsqueda aproximada de vecinos cercanos, lo que hacía que pareciera que no era posible cargar el índice. Finalmente, tras investigar este error y basándonos en las indicaciones dadas por el profesor y en el guion, decidimos optar por no usarlo y marcar esta opción como falsa, ya que al ser una colección pequeña no lo consideramos oportuno.

El segundo error vino al crear el índice con el uso de GPU, algo que por alguna razón no estaba habilitado o no funcionaba en el PC en el que se realizó, por problemas relacionados con los núcleos CUDA de la tarjeta gráfica, algo que iba a llevar a una configuración no solo del proyecto, sino también del sistema, de manera que se decidió prescindir de esta opción también.

Decisiones tomadas

Para la creación de los Run de cada índice, se han limitado los resultados de las búsquedas de cada *query* a 50 elementos.

Conclusión

Podemos ver que el mejor índice es el denso, ya que en general presenta unos resultados mayores para las métricas empleadas, las cuales son:

- *Hit Rate*, fracción de las queries para las cuales al menos un documento relevante es obtenido.
- *Precission*, proporción de los documentos obtenidos que son relevantes.
- *Recall*, porcentaje de entre los documentos relevantes que se han obtenido y los documentos relevantes.
- *MAP*, media de las puntuaciones de precisión calculadas después de recuperar cada documento pertinente.
- *MRR*, inverso multiplicativo del rango del primer documento pertinente recuperado: 1 para el primer puesto, 1/2 para el segundo, 1/3 para el tercero, etc...

Salida obtenida

```
COMPARACIÓN DE ÍNDICES DISPERSOS Y DENSOS
        Creando Qrels...
----
100%|██████████| 112/112 [00:00<00:00, 112034.83it/s]
  0%|          | 0/112 [00:00<?, ?it/s]
Ya existe el índice "cisi_sparse" así que lo cargamos.

        Creando Run: SparseRetriever...
100%|██████████| 112/112 [00:00<00:00, 301.24it/s]
----
Ya existe el índice "cisi_dense" así que lo cargamos.

        Creando Run: DenseRetriever...
100%|██████████| 112/112 [00:02<00:00, 48.83it/s]
----
#   Model              Hit Rate@50   P@50   Recall@50   MAP@50   MRR@50
--- -------------------- ------------- ------ ----------- -------- --------
a   CISI-SparseRetriever        0.961  0.158       0.289    0.133    0.542
b   CISI-DenseRetriever         0.961  0.176       0.307    0.138    0.558
```