Thống kê paper

-Theo bài báo : "Optimizing Dense Retrieval Model Training with Hard Negatives" thì in batch negative không khác gì random sampling negatives.

-Cũng bài báo "Optimizing Dense Retrieval Model Training with Hard Negatives" chỉ ra nếu chọn hard negative quá khó chúng sẽ khiến mô hình tập trung quá nhiều vào chúng và bỏ qua các ví dụ dễ dàng hơn.

- Có hai cách lấy hard negative là static và dynamic

- Bài báo " Sparse, dense, and attentional representations for text retrieval" và

"Noise-contrastive estimation: a new estimation principle for unnormalized statistical models"

chỉ ra rằng static không giúp hoặc làm giảm hiệu suất so với pretrained

-Bài báo :'Dense Passage Retrieval for Open-Domain Question Answering' chỉ ra rằng in-batchnegativetraining lấy 1 hard negative từ bm25 và gold postive đưa ra kết quả tốt nhất.

-Bài báo ‘LEARNING TO RETRIEVE: HOW TO TRAIN A DENSE RETRIEVAL MODEL EFFECTIVELY AND EFFICIENTLY’ chỉ ra rằng bm25 làm model bị bias giảm đồng bộ. Tác giả chỉ ra rằng BM25 dựa vào sự trùng lặp từ khóa giữa truy vấn và tài liệu để xếp hạng các tài liệu. Khi các tài liệu này được sử dụng làm ví dụ tiêu cực, mô hình DR sẽ học cách tránh truy xuất những tài liệu có nhiều từ giống với truy vấn.

-Bài báo ‘LEARNING TO RETRIEVE: HOW TO TRAIN A DENSE RETRIEVAL MODEL EFFECTIVELY AND EFFICIENTLY’ chỉ ra rằng Bm25 và random là tệ nhất khi chọn hard negative

-Bài báo ‘ Product quantization for nearest neighbor search. IEEE transactions on pattern analysis and machine intelligence’ chỉ ra rằng the BM25 score is instructive for training neural retrievers.

-Bài báo ‘ Complementing lexical retrieval with semantic residual Embedding’ chỉ ra dùng bm25 tốt

-Bài báo’Passage-based BM25 Hard Negatives: A Simple and Effective Negative Sampling Strategy For Dense Retrieval’ chỉ ra rằng dùng bm25 finetuned cũng ko tốt