Information Retrieval

Information Retrieval sẽ có 3 loại full-text search, vector search và hybrid search

I. Full-text search

- Một thuật toán search khá phổ biến trong IR là BM25. Đây là một loại thuật toán được xây dựng để cải tiến cho TF-IDF, nhằm cải thiện khả năng đánh giá đô quan trọng của các từ trong tài liệu so với toàn bộ văn bản.

1. Tính toán TF

- TF sẽ đo lường tần số xuất hiện của một từ khóa trong tài liệu. Bằng cách tính (số lần xuất hiện) / (Tổng số từ khóa).
- Tuy nhiên để tránh sự phụ thuộc vào quá nhiều của tần số xuất hiện nên BM25 sẽ làm mềm sự phụ thuộc đó.

2. Tính toán IDF

- IDF tính toán mức độ quan trọng của từ khóa bằng cách lấy nghịch đảo của tần số xuất hiện của từ đó.

3. Tính toán điểm số BM25

- Sau khi tính toán được điểm thành phần của TF và IDF. ĐIểm số của BM25 được tính toán như sau:

$$BM25(D,Q) = \sum_{i=1}^{n} IDF(t_1) \times \frac{f_{t_i,D} \times K_1 + 1}{f_{t_i,D} + k_1 \times (1 - b + b \times \frac{|D|}{avg\_dl})}$$

- t_i là từ khoá thứ i trong truy vấn Q
- $f_{t_i,D}$ là tần suất xuất hiện của từ khoá t_i trong tài liệu D.
- $|D|$ là độ dài của tài liệu D.
- avg_dl là độ dài trung bình của các tài liệu trong tập văn bản
- k_1 và b là các tham số điều chỉnh.

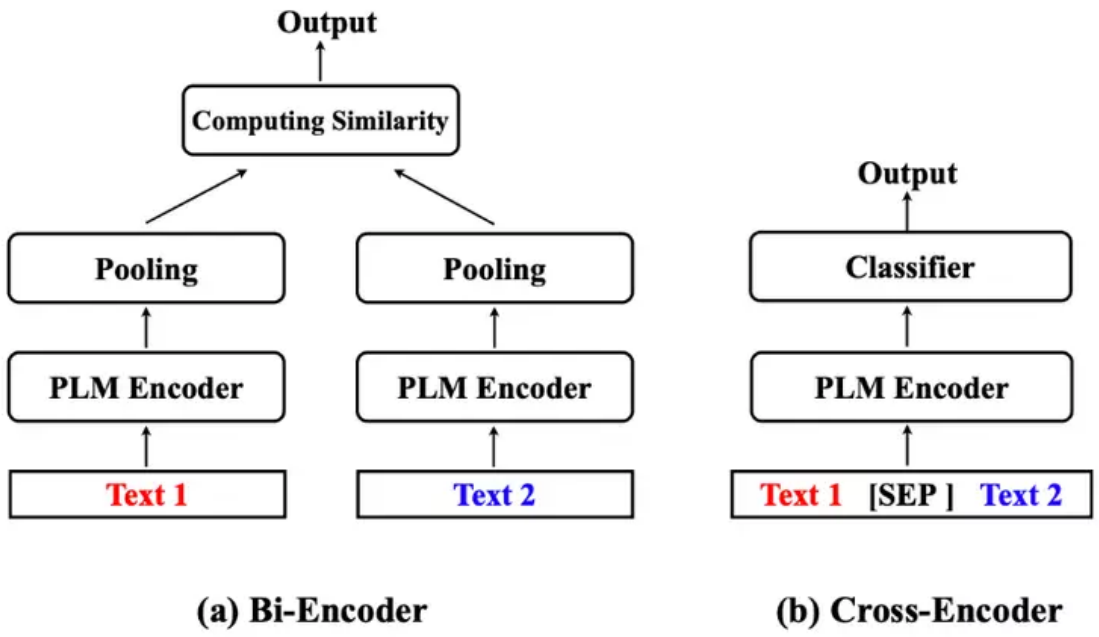
II. Vector Search

1. Bi-encoder

- Sử dụng một mô hình LLM để emb cho query và document. Sau đó sử dụng một số kỹ thuật tính độ tương đồng như cosin-similary, euclidean distance và jaccard similarity. Để tính ra được độ tương đồng để lấy top K cao nhất.

2. Cross-encoder

- Phương pháp này mang tính hiệu quả cao hơn là Bi-Encoder nhưng nó lại có nhược điểm là chậm hơn rất nhiều. Thông thường phương pháp này được sử dụng ở bước re-rank lại relevant documents sau khi được lấy ra từ Bi-encoder.
- Cách thực hiện của CE như sau:
  1. Đầu tiên nó sẽ combine lại câu query và document.
  2. Sau đó đi qua một Encoder để có thể tận dụng được self-attention của Encoder.
  3. Sau đó đi qua một lớp head classify để cho điểm score từ 0-1.



III. Hybrid Search

- Định nghĩa: Hybrid search là sự kết hợp giữa kết quả tìm kiếm của full-text search và vector search. Từ đó có thể kết hợp được ưu điểm của 2 loại này.

1. Hiệu xuất tìm kiếm

- Sparse retrieval (full-text search) thường được sử dụng để tìm kiếm theo từ khóa và trả về kết quả dựa trên độ tương đồng từ khóa.
- Dense retrieval (vector search) dùng để đánh giá độ tương đồng của câu query và documents.

2. Độ chính xác

- Sparse retrieval thường có độ chính xác cao trong việc đánh giá sự tương đồng của từ khóa và document.
- Dense retrieval có thể cung cấp một kết quả chính xác hơn

3. Đa dạng kết quả

- Khi kết hợp cả hai thì chúng ta sẽ có được nhiều tập dữ liệu được trả về giúp cung cấp thông tin đa dạng và phong phú hơn.

4. Tính linh hoạt

- Chúng ta có thể dễ dàng điều chỉnh số lượng trả về ở 2 loại trên.