

Neural inverted index for fast and effective information retrieval.

Information retrieval (IR) systems are designed to provide a ranked list of pertinent documents in response to user queries. Most contemporary information retrieval systems adopt the index-then-retrieve pipeline. Recently, an alternative approach called the Differentiable Search Index (DSI) has been proposed. Instead of segregating indexing and retrieval into two distinct components in an IR system, DSI aims to encompass all corpus information and execute retrieval within a single Transformer language model.

Task: The task is to build a model f that given a query q as input returns a ranked list of document ids. There should be a unified model trained to replicate the behavior of an index built on a corpus of documents and thereafter used to retrieve relevant documents. The proposed DSI should be different from the ones present in the literature.

Dataset: MS Marco. You can use the [Pyserini](#) library to work with it! Here you can find a colab notebook where you can see an [example](#) on how to use it.

Metrics:

- MAP: Mean Average Precision it's a crucial metric in the field of information retrieval. It's used to evaluate the effectiveness of search systems, like search engines or database queries. Here's a breakdown of what it means and how it's calculated:
 - Precision at K: Precision is a measure of relevancy. It's calculated as the number of relevant documents retrieved divided by the total number of documents retrieved. For example, if a search retrieves 10 documents and only 4 of them are relevant, the precision is 0.4.
 - Average Precision (AP): This is calculated for a single query. It's the average of the precision values calculated at the points in the ranking where each relevant document is retrieved. So, if there are several relevant documents, the precision is calculated each time one of these is encountered in the list of retrieved documents, and these values are averaged.
 - Mean Average Precision (MAP): This is the mean of the average precision scores for a set of queries. Essentially, you calculate the average precision for each query, and then find the mean of these values across all queries.
- Recall@1000: is the proportion of relevant items found in the top-1000 results.

References:

- <https://paperswithcode.com/paper/transformer-memory-as-a-differentiable-search>
- <https://arxiv.org/pdf/2305.02073.pdf>