Assignment 1

Studying effects of Workpiece Tokenization on Sparse Retrieval on a Corpus

Deadline : 5th September, 2023

You would have to explore the CISI dataset. In the following steps :

1. *Normal word-long tokenization* : Treat each word as a token. Index the corpus using the techniques covered in class (BM25, TF-IDF etc). You could use any of the many toolkits we would be suggesting in the upcoming sections.
2. *HuggingFace Workpiece Tokenization :* Originally designed to better capture subwords level dependencies to pre-train BERT. It is commonly used in many transformer based models to identify subword level patterns in dense datasets (unlike ours!). Tokenize the corpus using HuggingFace's Wordpiece Tokensizer and then re-run the same indexing system used above and compare the results.

You would be comparing performance based on two key metrics : MAP and NDCG . If you could come up with some other *sound and reasonable* metric of your own making feel free to compare using that as well.

Once you're done with this, you progress to the next step:

.

3. *Proximity Rewards :* Recall we covered Proximity Rewards in class, remark that the IR system you would be using in the previous two steps won't have any proximity rewards as such, so as a third step you need to add that to your IR System and again compare normal tokenization and Wordpiece based tokenization based on said metrics.

Submission Format (Subject to change?!)

We require you to do all your work on a Google Collab notebook after uploading the dataset to a preferred destination of the workspace and keeping its link in the very first cell of your collab notebook.

Suggested IR Systems:
Lemur from CMU
Whoosh
Elasticsearch
(Feel free to suggest more here, using teams as a discussion forum!)