SPLADE

Monique Monteiro – moniquelouise@gmail.com

Conceitos Importantes

Representação esparsa

- Propriedades desejáveis de modelos *bag-of-words* (*exact matching*, eficiência de índices invertidos, interpretabilidade)
- Similaridade semântica (expansão de *queries* e documentos)

Necessidade de métodos onde:

- A maior parte da computação possa ser feita offline
- Inferência online seja rápida

Sparse Lexical AnD Expansion (SPLADE)

- Expansão de documentos

Resultados obtidos

NDCG@10 = 0,7269 (vs BM25: 0,641)

Modelo utilizado: naver/splade-cocondenser-ensembledistil

Função de agregação: max (SPLADE v2)

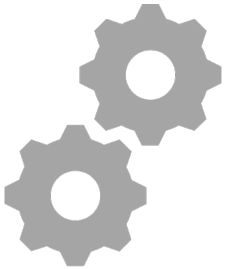
Experimentos realizados:

- Consulta direta à matriz de documentos em memória:
  - 0,393 segundos por query, 17 segundos para as 50 queries
  - Mais adequado para coleções pequenas (multiplicação de matrizes esparsas em GPU)
- Consulta a índice invertido:
  - 0,807 segundos por query, 42 segundos para as 50 queries
  - Escalável para coleções grande

Problemas e soluções encontrados

Uso da *attention mask*

Lembrar de utilizá-la na entrada e na saída



Oportunidades de otimização da construção do índice

Atualmente:

- Geração da matriz de documentos
- Construção do índice invertido a partir da matriz de documentos

Possibilidade de melhoria:

- Construção direta do índice invertido