TRESPI

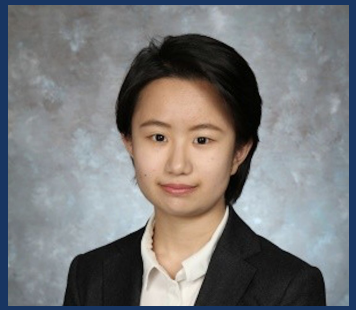
TRansformers for Expansion of SParse Indexes

A Research Project in Information Retrieval

Aug 3rd 2021

Berkeley
UNIVERSITY OF CALIFORNIA

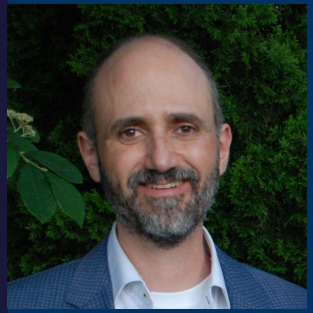
The “Bag of Words” team



Joanna Wang

Wade Holmes

Manpreet Khural



Stacy Irwin

Berkeley
UNIVERSITY OF CALIFORNIA

Introduction and Application of Information Retrieval

Current State:

- Strong general web search using features other than just webpage text

Practical Question:

- Can search be improved?
- Help institution find documents faster by returning best document in top position of results on the basis of document text content?

Research Motivation:

- HDCT Framework paper by Zhuyun Dai and Jamie Callan

Berkeley
UNIVERSITY OF CALIFORNIA

Common Terms in Information Retrieval

Ad hoc Information Retrieval	Retrieval of documents based on how well they match a query ● Example: Google search
Document Index	Inverted index used by document retrieval systems ● Index Key: term that appears in query ● Value: List of term weights for each document that contains the term
Context	Document text that surrounds query term. ● Indicates term's relevance ● Modifies term's meaning
Vocabulary Mismatch	Occurs when: ● Query's terms relevant to document's topic but… ● Term does not appear in document ● Occurs when query contains synonyms or descriptive phrases ● Examples: ○ USA, America ○ Capitol, Congress ○ Wall Street, Finance Sector

Reference:
Lin et. al., 2020

Berkeley
UNIVERSITY OF CALIFORNIA

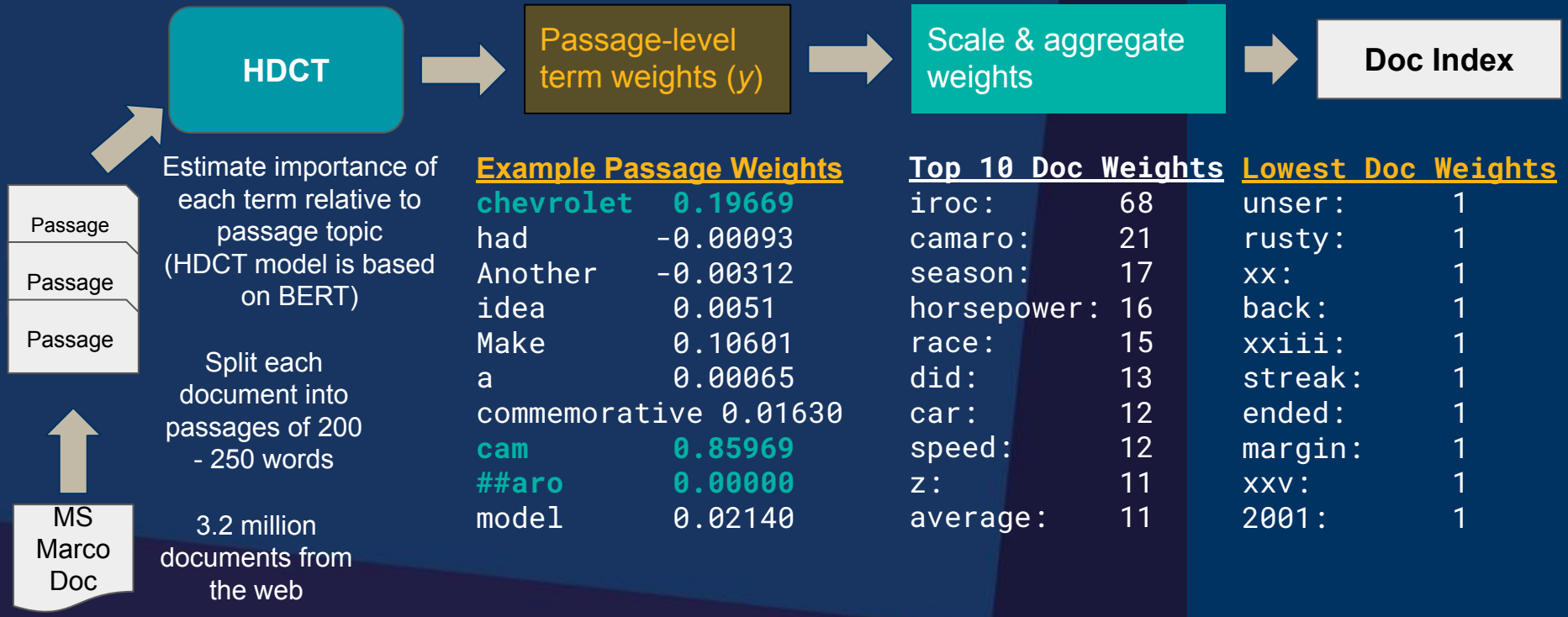
Starting Point: HDCT

- Focus Area: Information Retrieval – Document Retrieval
  - For an input query, output rank of most relevant documents
- Problem: How to improve over standard TF-IDF based Probabilistic IR?
- Solution: Calculate term weights using context (via BERT) rather than simple frequency
- Advantages:
  - Light on resources; uses existing Retrieval Algorithms (BM25)
  - Much more performant than complex, alternatives
- How have others built on this framework and how can we improve it?

Berkeley
UNIVERSITY OF CALIFORNIA

References:
Dai, Callan, 2020
Dai, Callan, 2019
Devlin et. al., 2019

HDCT Model - Generates Inverted Index



Source of example terms:

- MSMARCO Document D15
- http://www.iroczone.com/about/iroc-z-history/

Berkeley
UNIVERSITY OF CALIFORNIA

HDCT Retrieves Documents with BM25

Documents ranked against query with BM25 algorithm

- Probabilistic algorithm
- Fast, widely used, first implemented in 1980s

Calculates document-term weight for all terms that appear both in the query and document. Term weight inputs include:

- Term Frequency (TF)
  - Normally number of occurrences of term within document
  - HDCT replaces TF with HDCT term weight based on term importance
- Inverse Document Frequency (IDF)
- Document Length
- Average Document Length

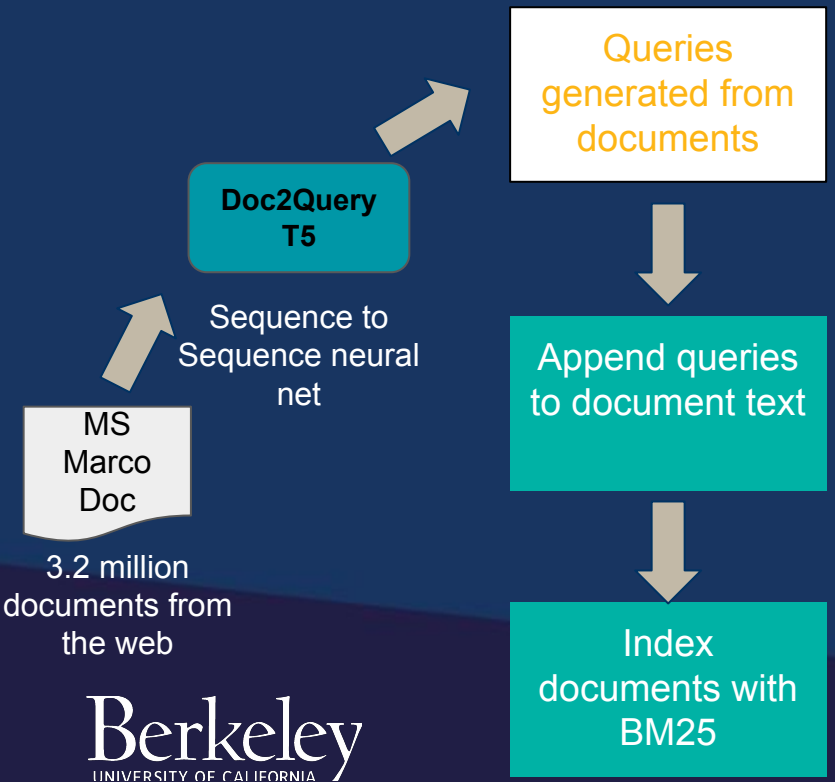
Document rank is sum of document-term weights

Query	BM25 doc score for D1573673
iroc camaro	19.7
iroc	12.0
porsche in nascar	5.1
ford pinto	0.0

Reference:
Robertson, Zaragoza, 2009

Berkeley
UNIVERSITY OF CALIFORNIA

DocT5Query Model



Queries

```
What car did iroc drivers drive?
Where is the first iroc race in nascar?
Which cars were the cars in the iroc?
…
```

190 generated queries contain 68 terms that do not appear in original document

New terms:

become, kind, buy, icoc, greatest, put, er, color, z, held, hq, newest, go, tpis, lights, makes, released, engines, size, icc, irc, icac, 's, v8, 1987, roc, market, imoc, hp, drove, chrome, win, irock, carries, opel, onele, between, rim, inoc, imorc-z, irocz, inverter, irac, original, irotc, icro, intercontinental, tecno, jeep, biggest

References:
Nogueira, Yang, et. al., 2019
Nogueira, Lin, et. al., 2019

Berkeley
UNIVERSITY OF CALIFORNIA

Research Question and Importance

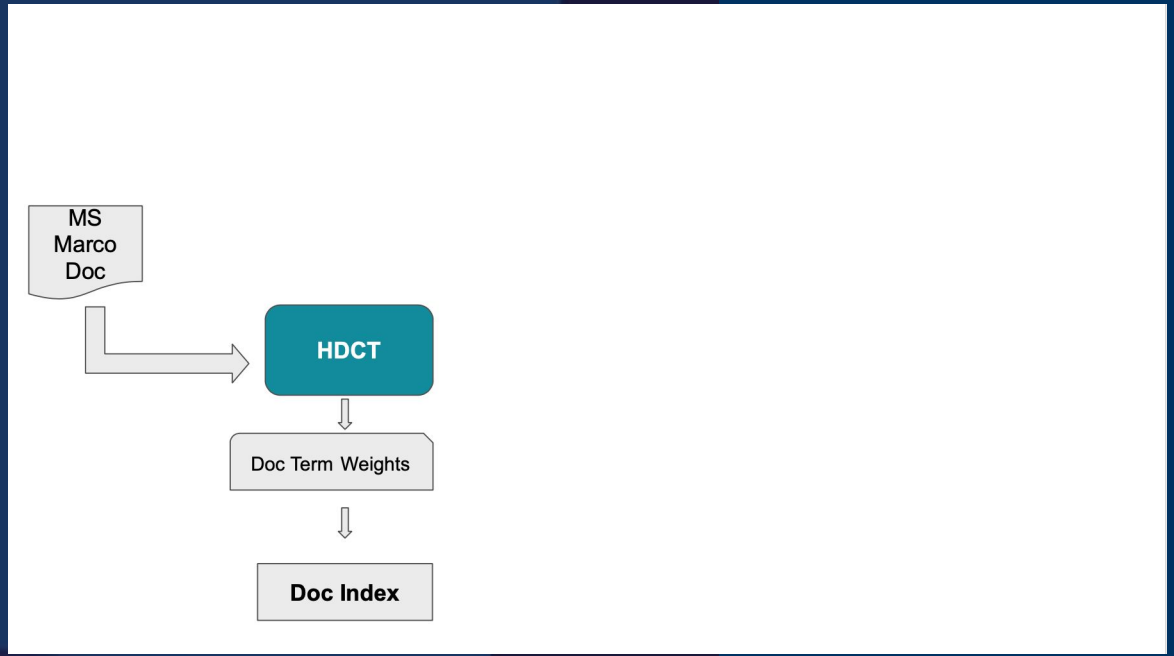
- HDCT: (+) Term Context (–) Vocabulary Mismatch
- DocT5Query:(+) Vocabulary expansion (–) No context
- **Research Question:** Can document indexes for information retrieval incorporate context AND be supplemented to address vocabulary mismatch as measured by document return rank?
- **Importance of work:** Index-based retrieval is core to even the most advanced multi-stage retrieval systems today. Small improvements in these systems have a material effect on the technology experience and performance.

Berkeley
UNIVERSITY OF CALIFORNIA

References:
Dai, Callan, 2020
Dai, Callan, 2019
Devlin et. al., 2019

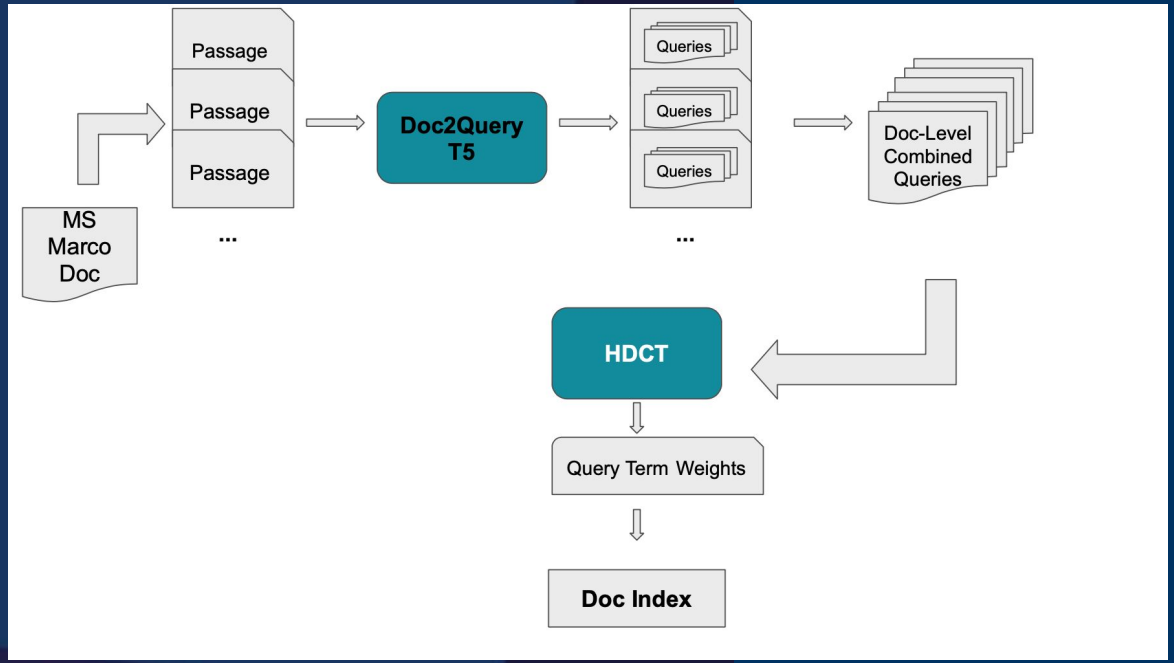
Solution: Generate and Combine Two Different Indexes

1. Generate index from source documents with HDCT



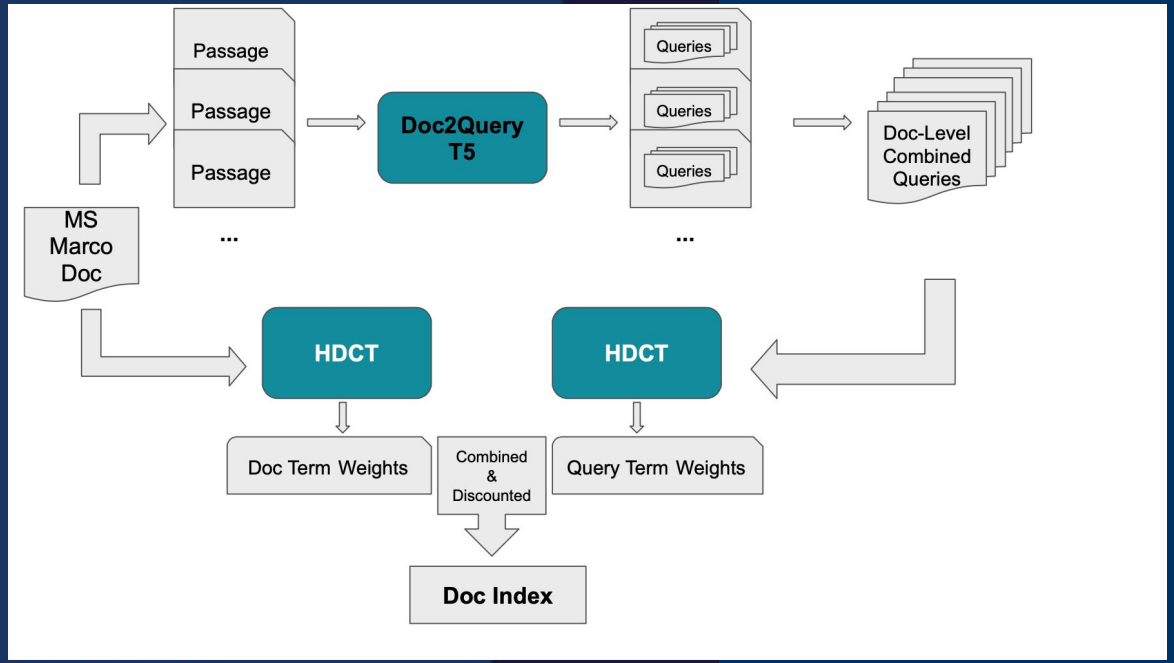
Solution: Generate and Combine Two Different Indexes

1. Generate index from source documents with HDCT
2. Generate index from queries with HDCT
   a. Generate queries from docT5query
   b. Combine queries into set of query documents
   c. Generate index with HDCT from query documents



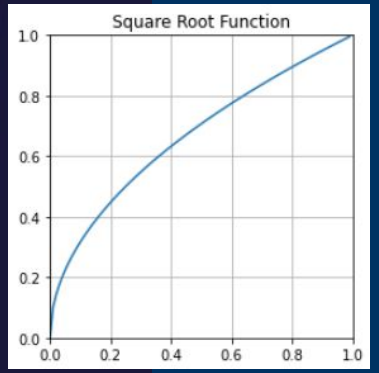
Solution: Generate and Combine Two Different Indexes

1. Generate index from source documents with HDCT
2. Generate index from queries with HDCT
   a. Generate queries from docT5query
   b. Combine queries into set of query documents
   c. Generate index with HDCT from query documents
3. Merge source document and query indexes
   a. Test different methods

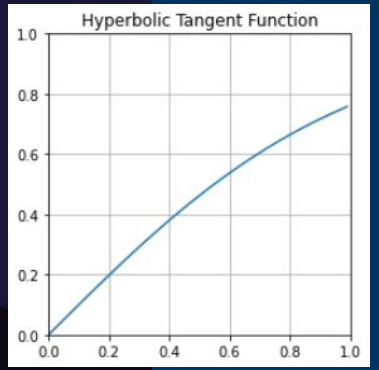


Passage Normalization Approach in TRESPI

- For smoothing, HDCT uses square root of initial term weight (*y*)
  - `passage_weight = round(sqrt(y) * 100)`
  - Raises low weights in comparison to higher weights
- TRESPI replaces square root with hyperbolic tangent
  - `passage_weight = round(tanh(y) * 100)`
  - Reduces the weight ratio for original vs new terms from 17:1 to 7.4:1
  - Improves MRR@100 and reduces index total size



$f(x) = x^{1/2}$



$f(x) = \tanh(x)$

Berkeley
UNIVERSITY OF CALIFORNIA

MS MARCO - The **Microsoft Machine Reading Comprehension Dataset**

"MS MARCO is a large scale dataset focused on machine reading comprehension, question answering, and passage ranking, Keyphrase Extraction, and Conversational Search Studies, or what the community thinks would be useful." - Microsoft

Dataset Details

- Dataset size: 3,213,835 documents queries
- Train subset: 367,013 Bing queries
- Dev: 5,193 Bing queries
- Train and Dev queries human verified and mapped to most relevant document id
- Test dataset is held private to discourage engineered performance, size unknown

Query	Relevant Document ID	Relevant Document Title
What are the two essential constituent elements of plain carbon steel?	D1862900	Difference Between Alloy Steel and Carbon Steel

Berkeley
UNIVERSITY OF CALIFORNIA

Demo

Berkeley
UNIVERSITY OF CALIFORNIA

MRR as evaluation criteria

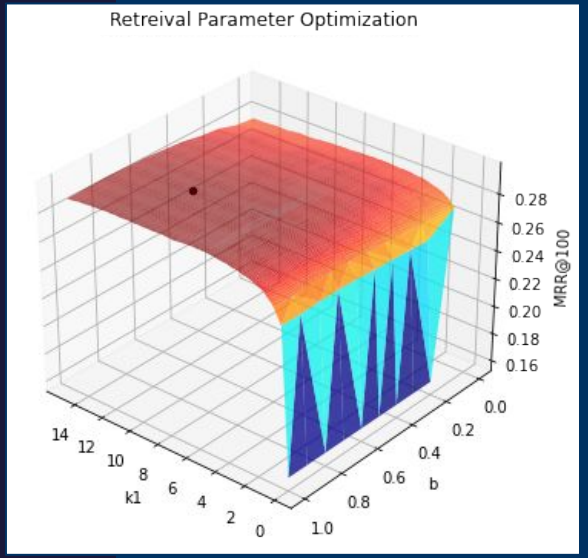
- **MRR is the right choice when:**
  - Order of the result matters
  - Only one document is expected to be the best result
  - The evaluation requirements are externally set
- **Other evaluation criteria may be better wehen:**
  - Multiple equally relevant documents are possible
  - The ability to find all documents is more important than one
  - Measuring performance of vocabulary mismatch models
  - Model recall is important to understanding performance
  - Large datasets where time performance is critical

Query	Results	Reciprocal Rank (RR)
What is smart infrastructure?	1. D1246238 2. D2845227 3. D2845225 4. D2845228 ...	RR = 1 / 3 = 0.3333
What are the two essential constituent elements of plain carbon steel?	1. D2897426 2. D2235377 3. D2156745 4. D2147907 ...	Document does not appear in top 10 results: RR = 0
How many calories do you have to burn to lose a pound?	1. D2466979 2. D1347781 3. D828735 4. D60587 ...	RR = 1 / 1 = 1.000 Note: spelling error (“loose”) occurred in both qrel and original document

Berkeley
UNIVERSITY OF CALIFORNIA

Model Performance: Results

Model Description	MRR@100	Delta	k1	b
Baseline HDCT	0.20889	-	0.9	0.4
T5 body only	0.26715	0.05825	0.9	0.4
Tanh passage normalization	0.24821	0.03932	0.9	0.4
T5 body and HDCT body	0.26604	0.05715	0.9	0.4
Averaging Method	0.25065	0.04176	0.9	0.4
Max Method	0.26151	0.05262	0.9	0.4
T5 body with tanh	0.27155	0.06266	0.9	0.4
T5 body with tanh (optimized)	**0.29251**	**0.08362**	**10.3**	**0.7**
Baseline HDCT (optimized)	0.27608	-	10	0.9



Note: These results are based on the passage averaging method, decay results available in the github repository results table.

Berkeley
UNIVERSITY OF CALIFORNIA

Wrap-Up

- We built on HDCT/DeepCT and docT5query to improve MRR@100 by 0.062 over baseline and 0.0836 with hyperparameter tuning
- The approach is practical and can be implemented using standard BM25 retrieval with a smaller index size than HDCT
- Check the MS Marco Document Retrieval leaderboard for final ranking

Recommendations

- Different evaluation metric; take into account more than just predicted rank of most relevant document
  - MRR vs Precision vs nDCG vs Spearman's Rank Correlation Coefficient
- Explore larger data sets: ClueWeb
- Train DocT5Query specifically for task of vocabulary expansion
  - Refined query generation to eliminate useless terms and increase synonym capture
- Neural network based term weight combination

Questions

Berkeley
UNIVERSITY OF CALIFORNIA

References

1. Z. Dai, J. Callan. 2019. *Context Aware Sentence/Passage Term Importance Estimation for First Stage Retrieval.* https://arxiv.org/abs/1910.10687
2. Z. Dai, J. Callan. 2020. *Context aware document term weighting for ad-hoc search.* In Proc. of The Web Conference 2020, 2020, pp. 1897–1907. https://dl.acm.org/doi/10.1145/3366423.3380258
3. J. Devlin, M. Chang, K. Lee, K. Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1. https://aclanthology.org/N19-1423.pdf
4. J. Lin, R. Nogueira, A. Yates. 2020. *Pretrained Transformers for Text Ranking: BERT and Beyond.* https://arxiv.org/pdf/2010.06467
5. R. Nogueira, W. Yang, J. Lin, K. Cho. 2019. *Document Expansion by Query Prediction.* https://arxiv.org/abs/1904.08375
6. R. Nogueira, J. Lin. 2019. *From doc2query to docTTTTTquery.* https://cs.uwaterloo.ca/~jimmylin/publications/Nogueira_Lin_2019_docTTTTTquery.pdf
7. S. Robertson, H. Zaragoza. 2009. *The Probabilistic Relevance Framework: BM25 and Beyond.* Foundations and Trends in Information Retrieval. Vol. 3, No. 4 (2009) pp. 333–389. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.437.660&rep=rep1&type=pdf

Berkeley
UNIVERSITY OF CALIFORNIA