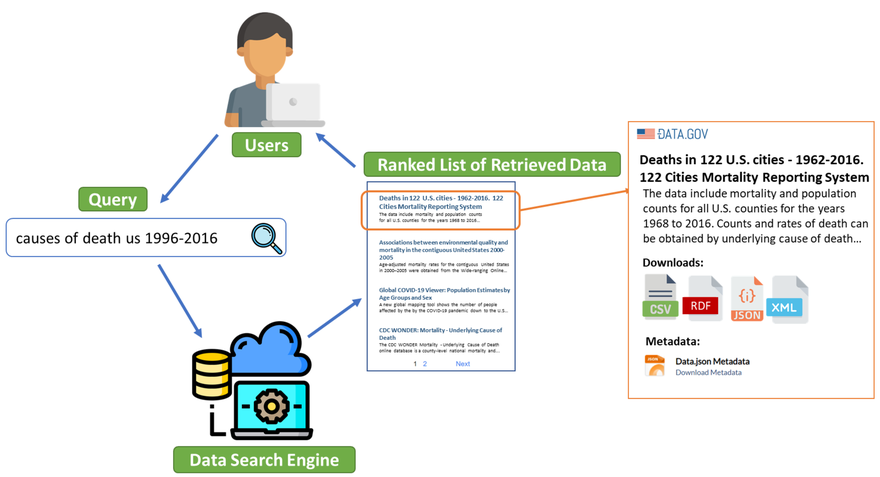
STIS at the NTCIR16-DataSearch 2 Task: Ad-hoc Data Retrieval Ranking with Pretrained Representative Words Prediction

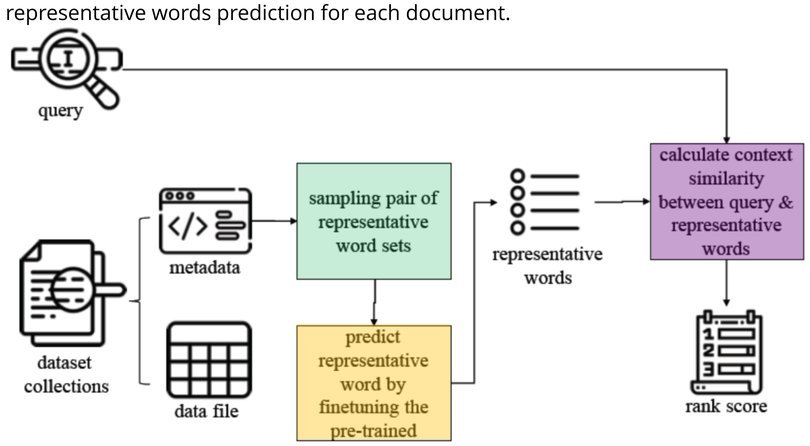
Lya Hulliyyatus Suadaa, Lutfi Rahmatuti Maghfiroh, Muhammad Luqman and Isfan Nur Fauzi
STIS Polytechnic of Statistics
Indonesia

Task Introduction



Proposed Ranking Architecture

Implemented **Pre-training with Representative wOrds Prediction (PROP)** for ad-hoc retrieval (Xinyun Ma, et al., WSDM 2021) in constructing the representative words prediction for each document.



Representative Word Sets Sampling

- sample a pair of representative word sets from vocabulary $V = \{w_i\}_n^1$ based on document language model following dirichlet distribution with probability P(wi|D)

$$\text{if } QL_u > QL_v \text{ then } w_{pos} = u \text{ and } w_{neg} = v$$
$$\text{else } w_{pos} = v \text{ and } w_{neg} = u$$

Representative Words Prediction

- finetune pre-trained Transformer BERT for representative words prediction task
- hidden state: $h_{[CLS]} = \text{Transformer}([CLS] + w_{rep} + [SEP] + D + [SEP])$ where $wrep = \{wpos, wneg\}$
- probability of word represent to the the document: $P(w_{rep}|D) = \text{MLP}(h_{[CLS]})$
- loss function: $\mathcal{L} = \max(0, 1 - P(w_{pos}|D) + P(w_{neg}|D))$

Rank Score Calculation

- rank score using prop: $S_{prop} = \text{avg}(BERT_{Score}(query, w_{prop_i})), i = 1...k$

  where $\{w_{prop}\}_1^k$: k list of representative word prediction from the finetuned BERT

Ranking by PROP and BertScore	Traditional IR + Reranking by PROP and BertScore
$S = S_{prop}$	$S = \alpha S_{base} + (1- \alpha)S_{prop}$

Results

Model	nDCG@3	nDCG@5	nDCG@10	nERR@3	nERR@5	nERR@10	Q-mea
BM25	**0.191**	**0.188**	**0.211**	**0.199**	**0.222**	**0.233**	**0.248**
Ranking by PROP and BertScore	0.172	0.175	0.201	0.191	0.188	0.201	0.218
BM25 + Reranking by PROP and BertScore	0.163	0.173	0.202	0.192	0.183	0.200	0.221

Table 1: Results from NTCIR-16 Data Search 2. The best score is in bold.

Query	Representative Word	$Rank_{BM25}$	$Rank_{PROP}$
annual turnover care workers	disabled, incoming, benefits statistics, insurance social, socially calculations workers, securing three, three, receiving edition calculate, disability series people disability tables workers, peoples social statistical	56	1
	low researchers, teams v u shorebird shores forage, forest v tennessee foraging group, radio forest primary radio migrate, ponds estimating migratory shorebird, opening vulnerable estimated network rates, calidris, western flight shorebird reservation created migration, establishing rate, build regional valleys	1	35

Table 2: Samples of representative words prediction for each query and their ranks.

DATA.GOV
Annual Statistical Report on the Social Security Disability Insurance Program, 2005
This annual report provides program and demographic information on the people who receive Social Security Disability Insurance Program benefits. This edition presents a series of detailed tables on the three categories of beneficiaries: **disabled workers, disabled widowers, and disabled adult children**. Numbers presented in these tables may differ slightly from other published statistics because all tables...
Downloads:

Representative Words:
disabled, incoming, benefits statistics, insurance social, socially calculations workers, securing three, three, receiving edition calculate, disability series people disability tables workers, peoples social statistical

DATA.GOV
Turnover Rates of Fall Migrating Pectoral Sandpipers Through the Lower Mississippi Alluvial Valley
The Mississippi Alluvial Valley (MA V) is the historic alluvial floodplain of the Lower Mississippi River.... The primary objective of this study is to estimate the turnover rates of fall migrating shorebirds in the MA V using 2 target species, the pectoral sandpiper (Calidris metanotos, PESA) and the least sandpiper (Calidris minulilla, LESA). We will estimate turnover rates from PESAs using radio telemetry data from 2001 and 2002 and from LESAs using capture-resight data from 2002.
Downloads:

Representative Words:
low researchers, teams v u shorebird shores forage, forest v tennessee foraging group, radio forest primary radio migrate, ponds estimating migratory shorebird, opening vulnerable estimated network rates, calidris, western flight shorebird reservation created migration, establishing rate, build regional valleys

Conclusion

Addressing ad-hoc retrieval approach for governmental statistical data:
- proposed using a pre-trained model to capture representative words prediction for each document then calculate the similarity between the query and the representative words as a rank score.
- proposed combined the representative similarity score to re-rank candidate documents of BM25 model for each query.