COMP0084 Coursework 1

Anonymous ACL submission

Abstract

None

1 Task 1: Evaluating Retrieval Quality

1.1 Text Processing

We preprocess passages/queries in two files: *validation_data.tsv* and *train_data.tsv* by following steps:

1. remove url.
2. lower characters.
3. remove non alpha characters.
4. tokenization by Python package NLTK.

1.2 BM25 Model

BM25 Model with parameters implemented in the Coursework 1 is used to retrieval the top passages for each query. Here we retrieval top 3, 10 and 100 passages from 1000 passages. The complete result of BM25 model is in *bm25_raw_top1000.tsv*, while we retrieval top 3 passages for each query in *bm25_ordered_top3.tsv*, top 10 passages in *bm25_ordered_top10.tsv* and top 100 passages in *bm25_ordered_top100.tsv*.

1.3 Metrics

1.3.1 Average Precision (AP)

AP is the average precision of relevant passages for a query.

$$AP = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{N} \quad (1)$$

N: number of relevant passages for the query.
k: rank of the passage.

1.3.2 Mean AP

Mean AP is the average of AP over all queries.

$$mAP = \frac{\sum_{q=1}^{N_q} AP_q}{N_q} \quad (2)$$

q: the q_{th} query

1.3.3 Discounter Cumulative Gain (DCG)

DCG is the total gain accumulated at a particular rank p.

$$DCG_q = \sum_{i=1}^{q} \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad (3)$$

p: particular ranking length
i: the ith passage
rel_i: relevance score

1.3.4 Normalized DCG (NDCG)

Normalizes DCG against the best possible DCG result (the perfect ranking) for a query.

$$NDCG_q = \frac{DCG_q}{optDCG_q} \quad (4)$$

1.3.5 Mean NDCG

Mean NDCG is the average of NDCG over all queries.

$$mNDCG = \frac{\sum_{q=1}^{N_q} NDCG_q}{N_q} \quad (5)$$

N_q: number of queries
The results of the metrics of BM25 model are in Table 1.

Table 1: Metrics of BM25 model

BM25	Cutoff	mAP	mNDCG
Top 3	3	0.1830	0.2007
Top 10	10	0.2250	0.2870
Top 100	100	0.2367	0.3548

2 Task 2: Logistic Regression

3 Task 3: LambdaMART Model