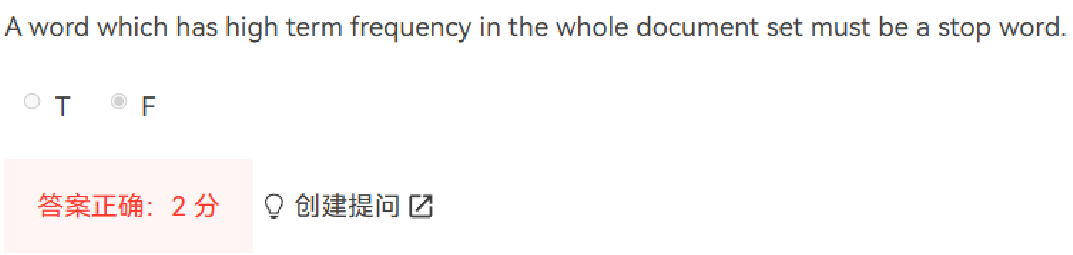
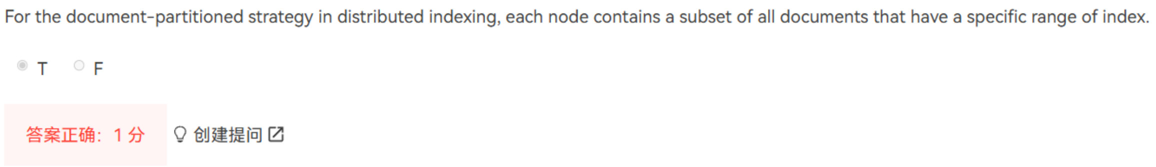


A word which has high term frequency in the whole document set must be a stop word.

○ T ● F

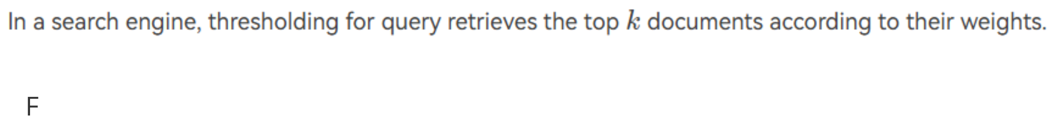
答案正确：2 分 创建提问



For the document-partitioned strategy in distributed indexing, each node contains a subset of all documents that have a specific range of index.

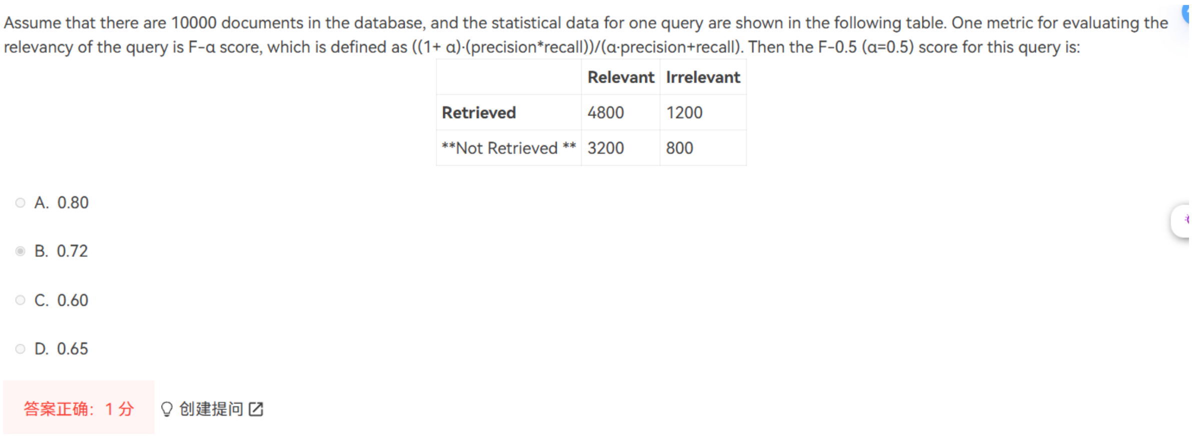
● T ○ F

答案正确：1 分 创建提问



In a search engine, thresholding for query retrieves the top k documents according to their weights.

F



Assume that there are 10000 documents in the database, and the statistical data for one query are shown in the following table. One metric for evaluating the relevancy of the query is F-α score, which is defined as ((1+ α)·(precision*recall))/(α·precision+recall). Then the F-0.5 (α=0.5) score for this query is:

	Relevant	Irrelevant
Retrieved	4800	1200
**Not Retrieved **	3200	800

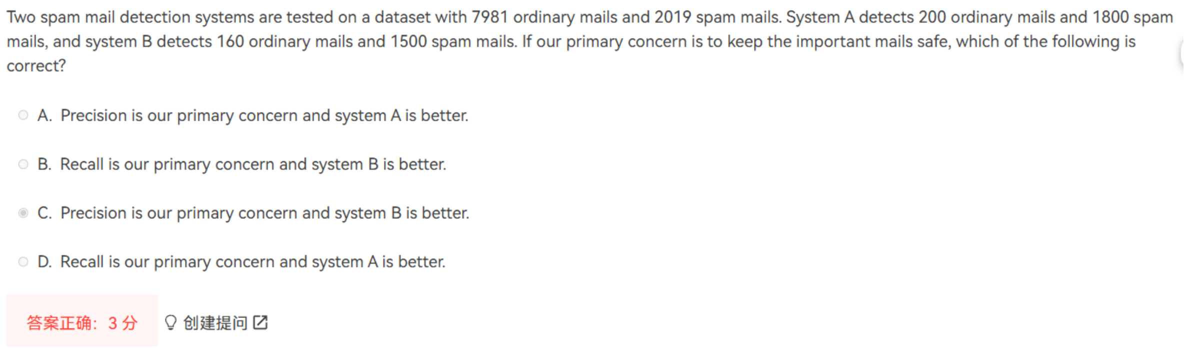
○ A. 0.80

● B. 0.72

○ C. 0.60

○ D. 0.65

答案正确：1 分 创建提问



Two spam mail detection systems are tested on a dataset with 7981 ordinary mails and 2019 spam mails. System A detects 200 ordinary mails and 1800 spam mails, and system B detects 160 ordinary mails and 1500 spam mails. If our primary concern is to keep the important mails safe, which of the following is correct?

○ A. Precision is our primary concern and system A is better.

○ B. Recall is our primary concern and system B is better.

● C. Precision is our primary concern and system B is better.

○ D. Recall is our primary concern and system A is better.

答案正确：3 分 创建提问