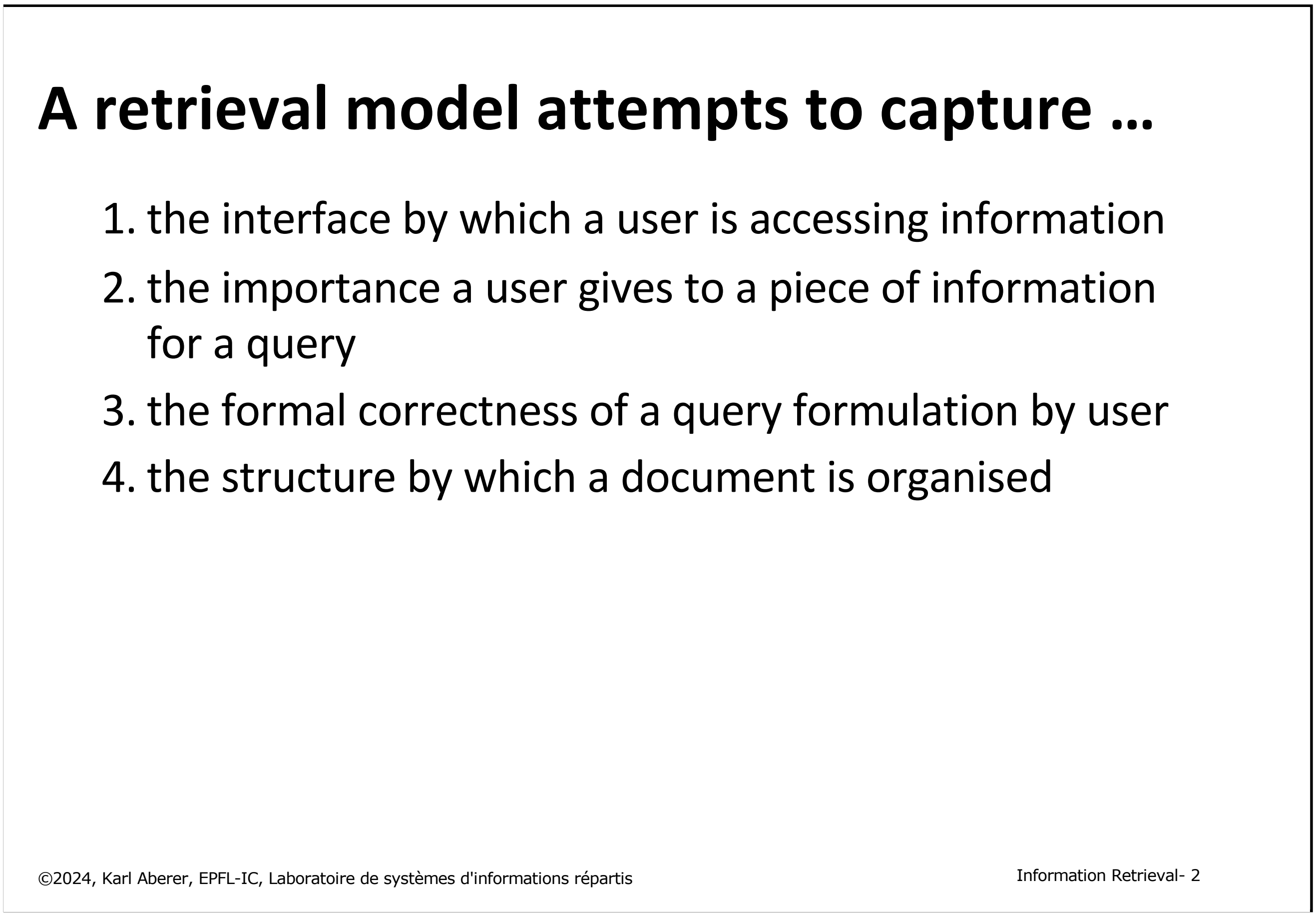
Answers to quizzes

1.1 INTRODUCTION TO INFORMATION RETRIEVAL

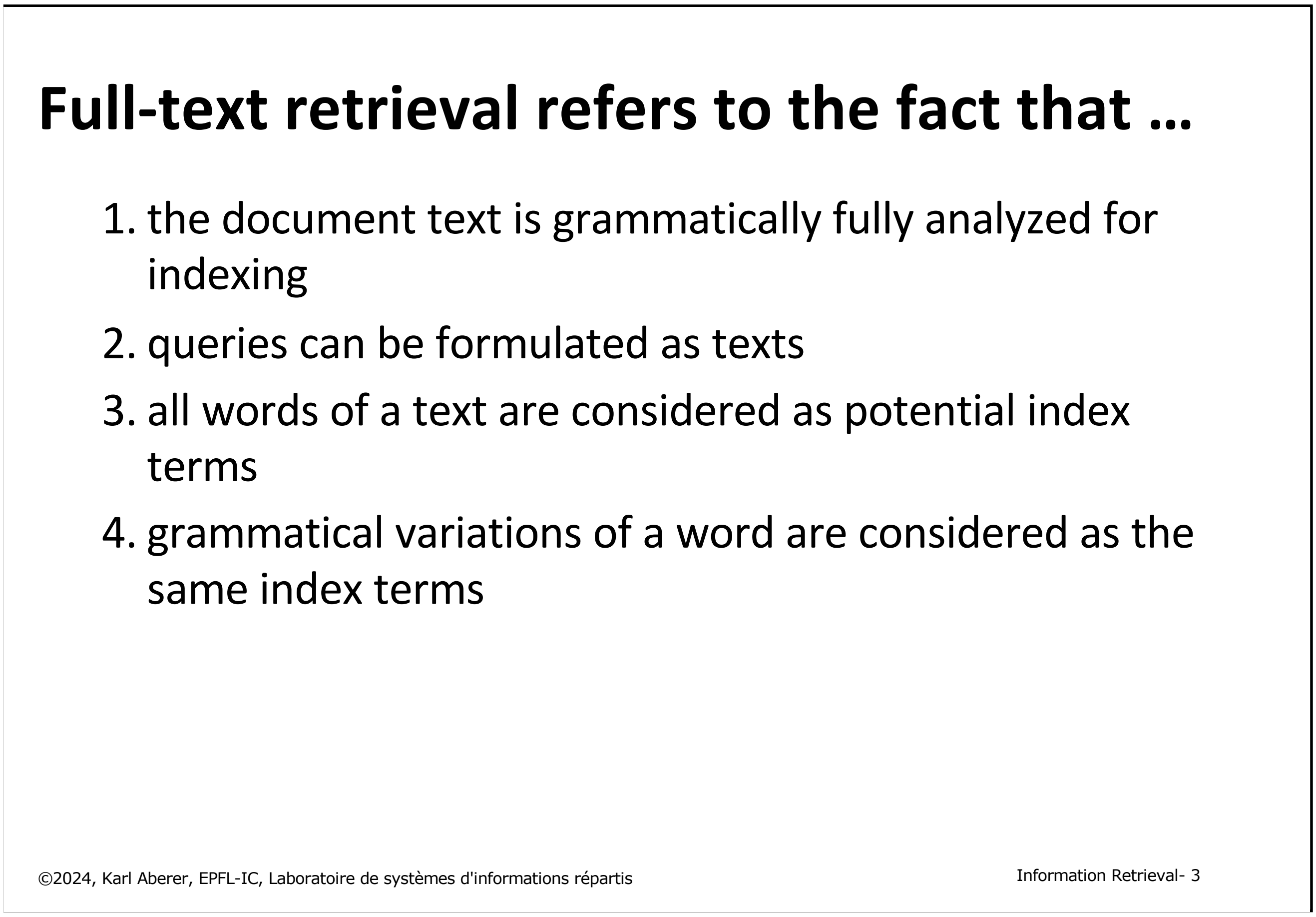


A retrieval model attempts to capture …

1. the interface by which a user is accessing information
2. the importance a user gives to a piece of information for a query
3. the formal correctness of a query formulation by user
4. the structure by which a document is organised

Answer 2

The role of a retrieval model is to capture the notion of relevance, which means what documents a user would consider as relevant for a given query. The user interface is an orthogonal question. If the retrieval model involves a notion of formal correctness of queries, it is also verified in a user-interfacing application. The structure of a document can be considered by a retrieval model, but is not modelled by it.

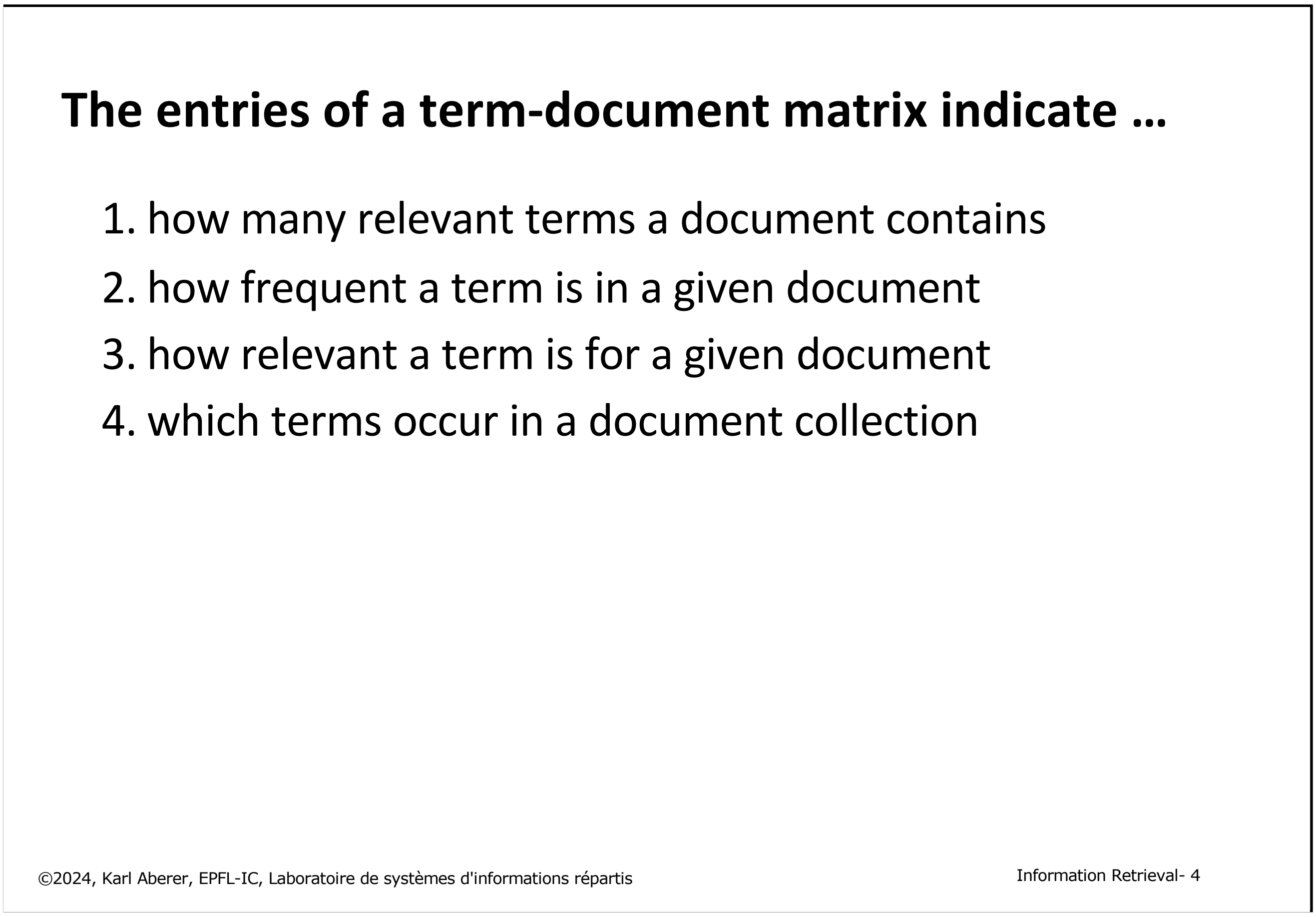


Full-text retrieval refers to the fact that …

1. the document text is grammatically fully analyzed for indexing
2. queries can be formulated as texts
3. all words of a text are considered as potential index terms
4. grammatical variations of a word are considered as the same index terms

Answer 3

The term full-text retrieval has been introduced to the fact that the text is considered as a bag of words. This means that any grammatical structure us ignored. Considerung grammatical variations of a word as the same index term is achieved by using stemming.

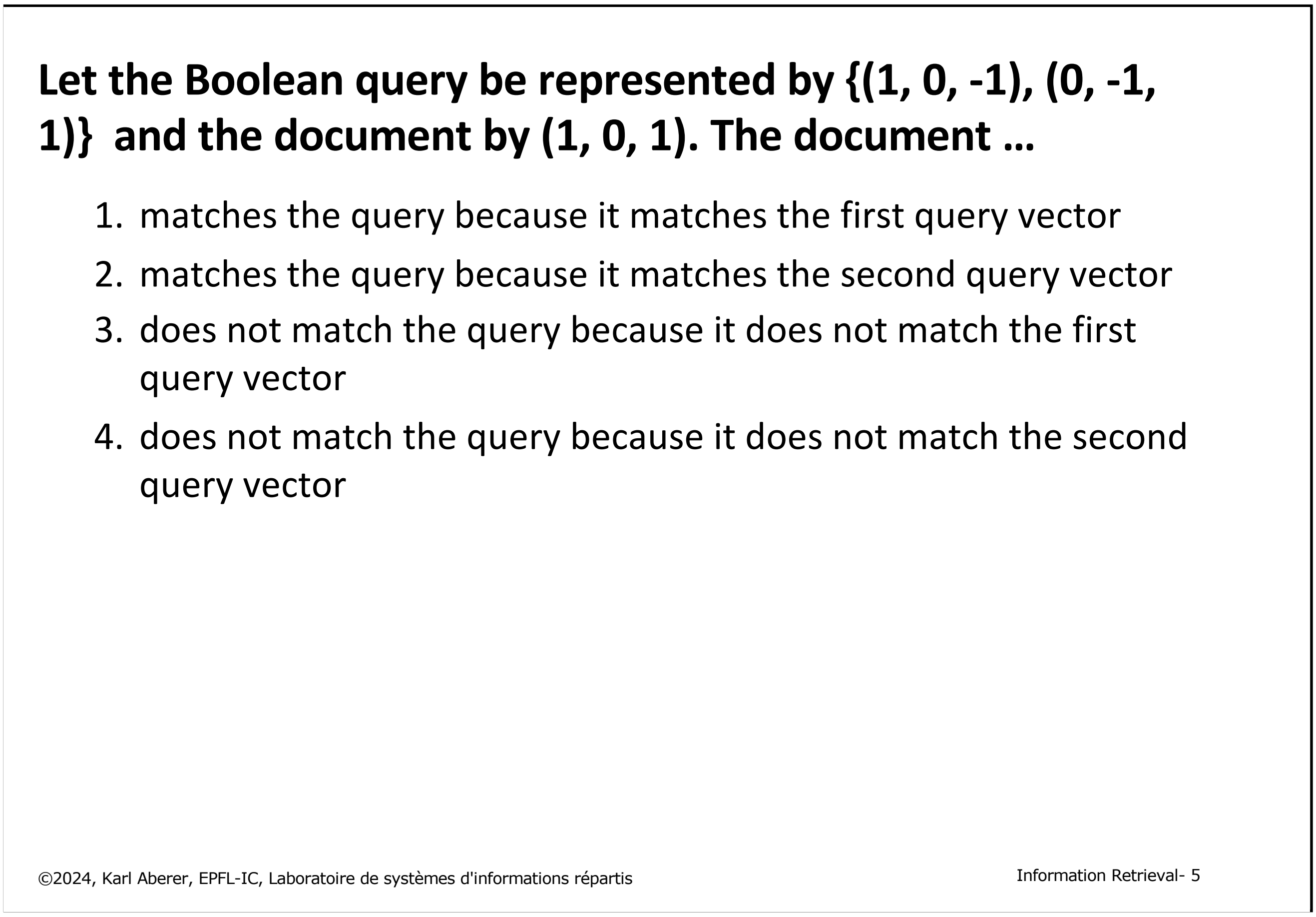


The entries of a term-document matrix indicate …

1. how many relevant terms a document contains
2. how frequent a term is in a given document
3. how relevant a term is for a given document
4. which terms occur in a document collection

Answer 3

As per definition the entries determine the relevance of a term to the document. The frequency can be such an indicator, but is not necessarily so. An entry in the matrix does not indicate which terms occur an a document collection, but whether a term occurs (otherwise the entry would be a set). It does also not indicate how many relevant terms a document contains, sicne the entry is linked to a single term.

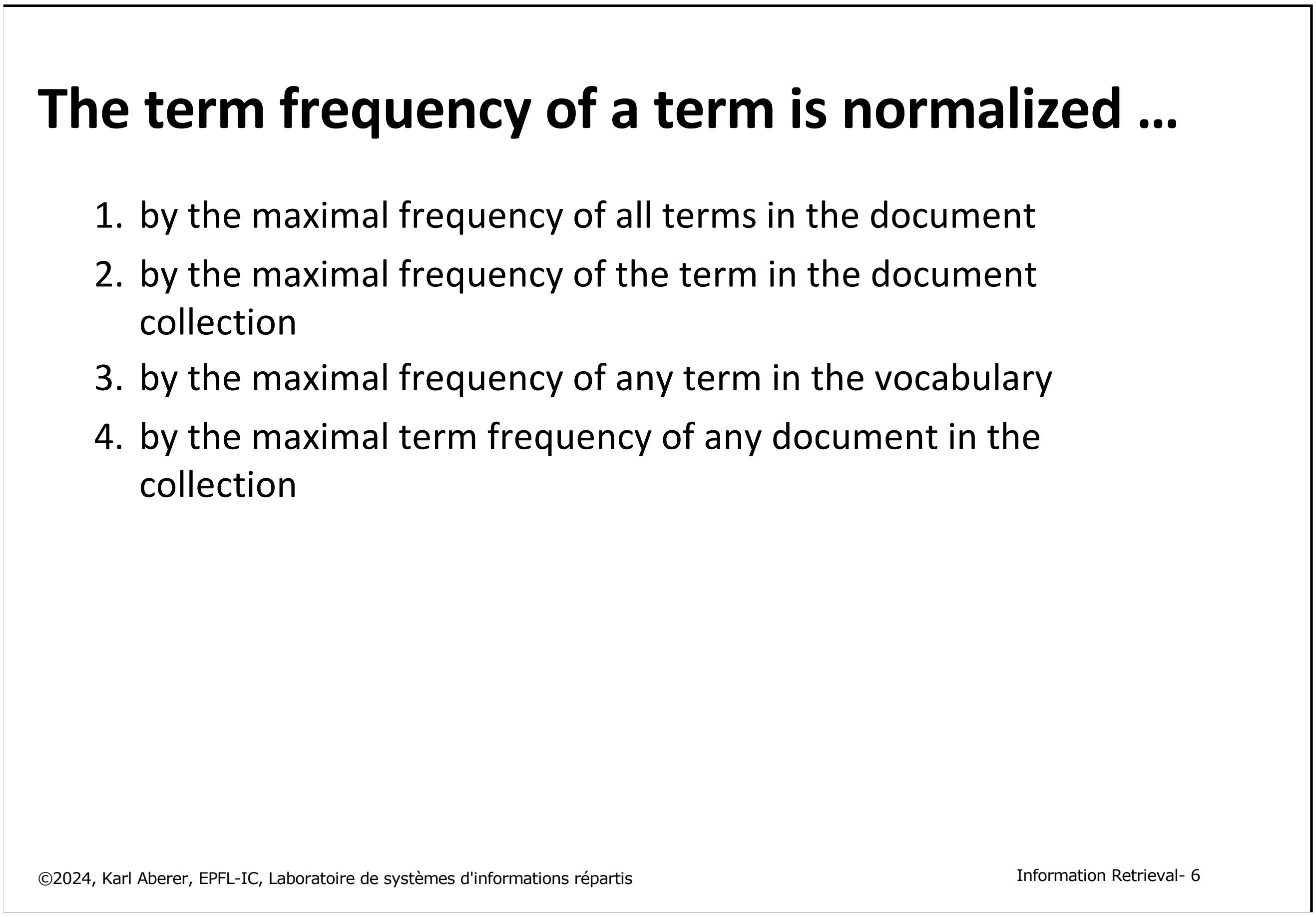


Let the Boolean query be represented by {(1, 0, -1), (0, -1, 1)} and the document by (1, 0, 1). The document …

1. matches the query because it matches the first query vector
2. matches the query because it matches the second query vector
3. does not match the query because it does not match the first query vector
4. does not match the query because it does not match the second query vector

Answer 2

The document vector (1,0,1) matches the query vector (0,-1,1), since the first word is not relevant for the query, the second word has to be absent from the document, which is the case since the value is 0, and the third word has to be present in the document, which is the case since the value is 1.

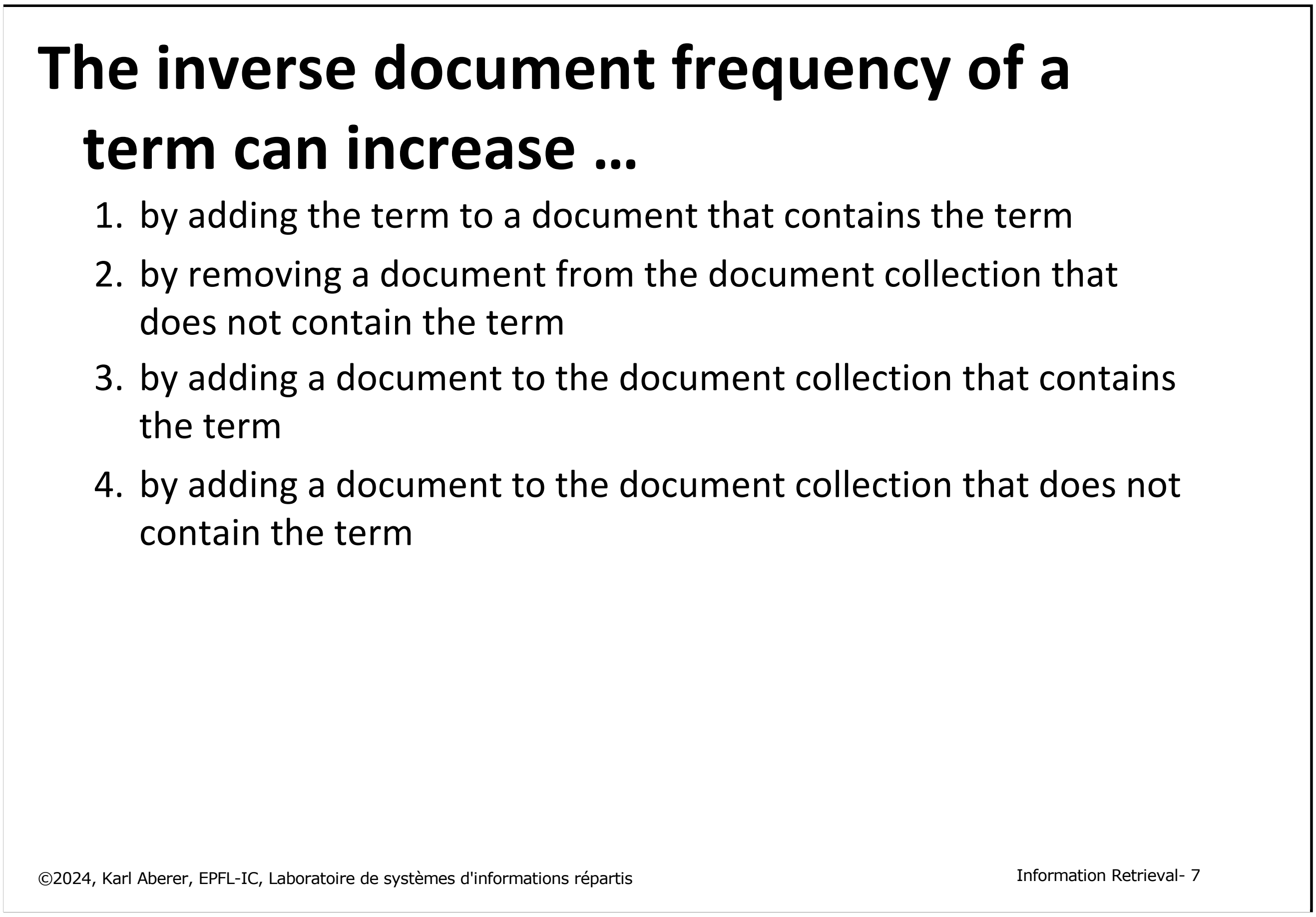


The term frequency of a term is normalized …

1. by the maximal frequency of all terms in the document
2. by the maximal frequency of the term in the document collection
3. by the maximal frequency of any term in the vocabulary
4. by the maximal term frequency of any document in the collection

Answer 1

The standard normalization of term frequency is by the maximal frequency of all terms occuring in the document. Note the denominator in the definition, which expresses that you take for all words k in the vocabulary the frequency of the word in the document j. This implies that the most frequent word will have normalized term frequency 1. It implictly also takes into account the document length, since in a longer document the most frequent words will be more frequent and this the influence of a single word will be smaller.

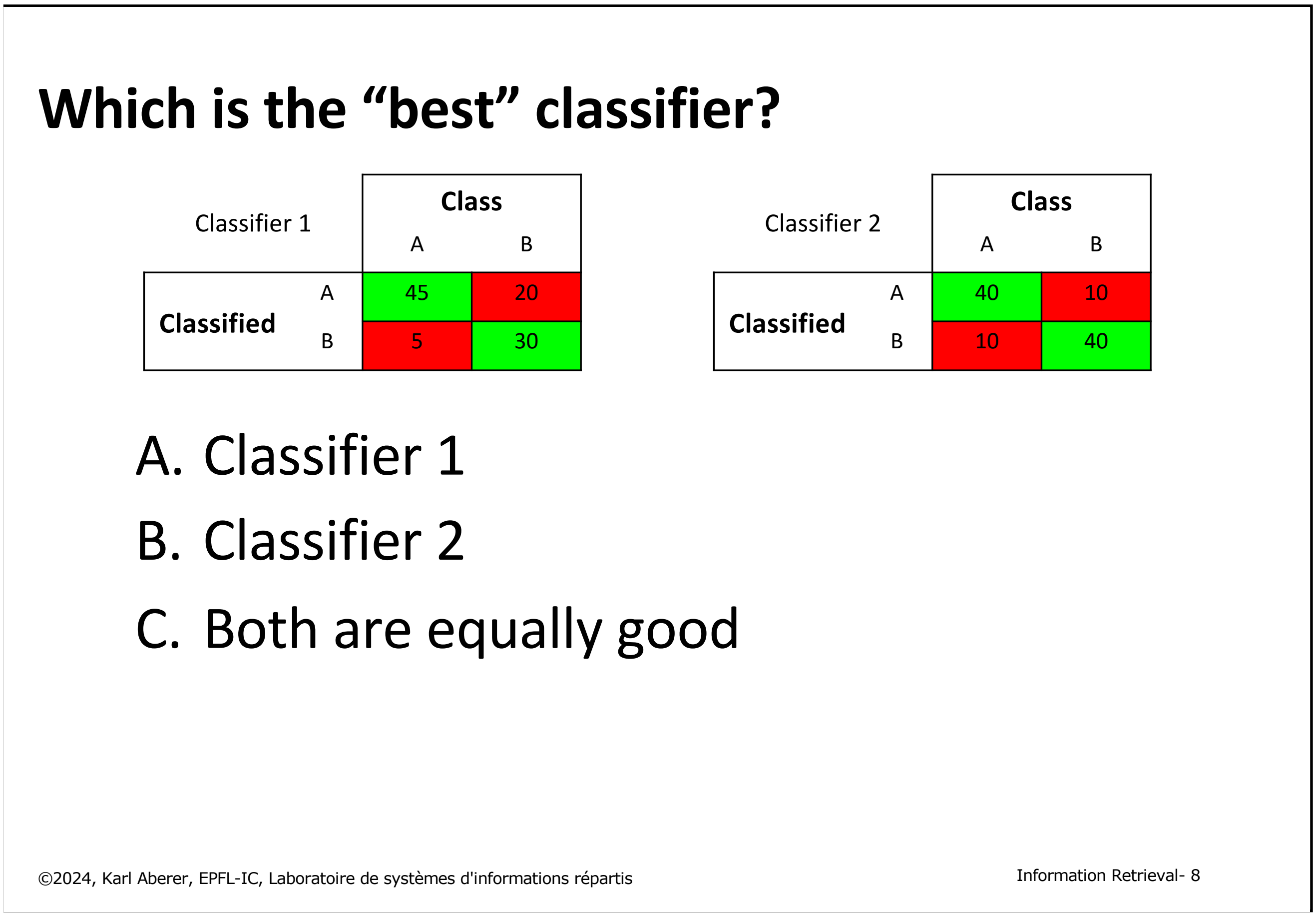


The inverse document frequency of a term can increase …

1. by adding the term to a document that contains the term
2. by removing a document from the document collection that does not contain the term
3. by adding a document to the document collection that contains the term
4. by adding a document to the document collection that does not contain the term

Answer 4

The inverse document frequency of a term increases if the proportion of documents containing the term decreases. By adding a document to the collection that does not contain the term this is the case. In the first case the idf remains unchanged, in the second and thrid case it decreases.



Which is the “best” classifier?

Classifier 1		Class	
		A	B
Classified	A	45	20
	B	5	30

Classifier 2		Class	
		A	B
Classified	A	40	10
	B	10	40

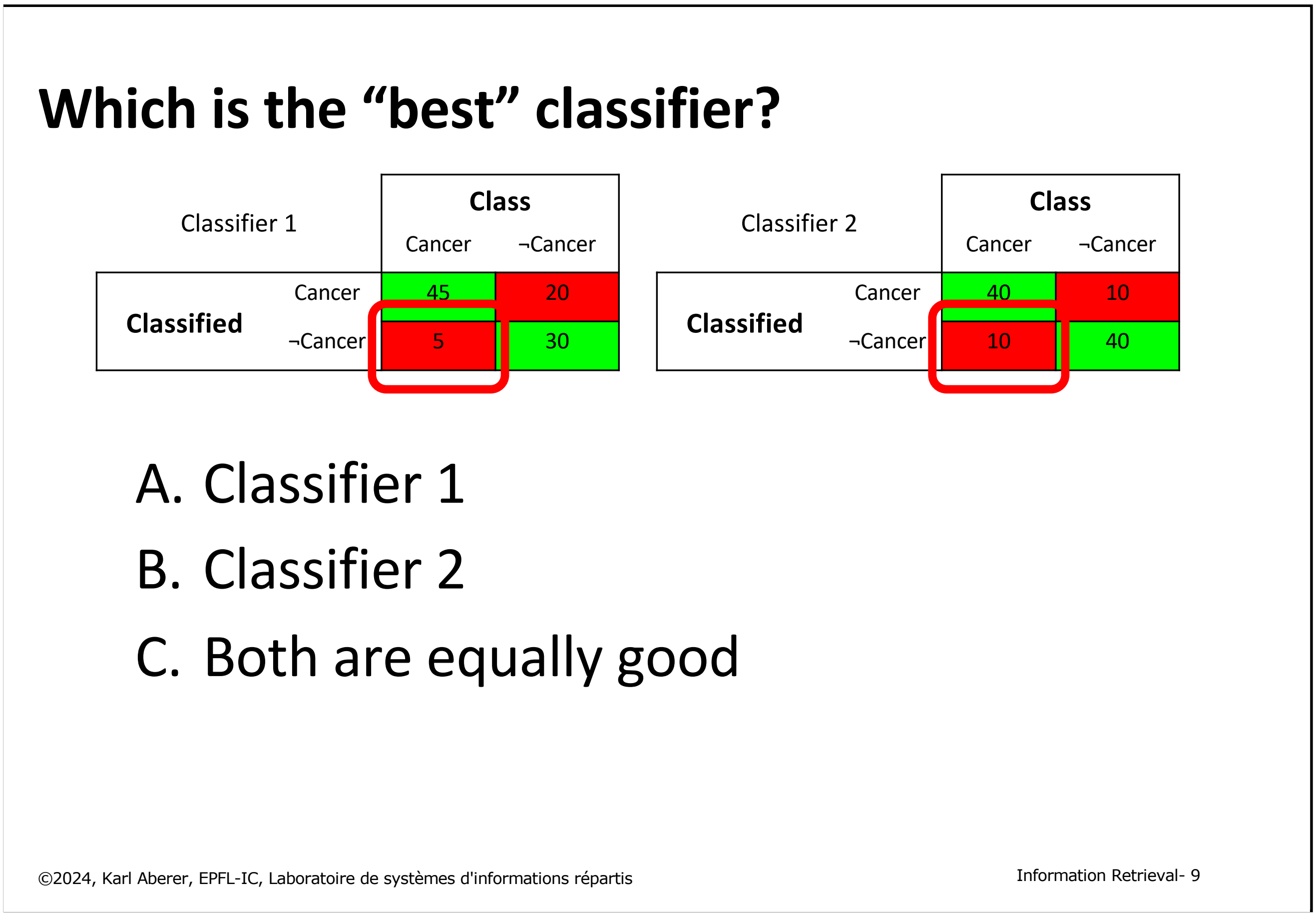
A. Classifier 1

B. Classifier 2

C. Both are equally good

Answer B

Without any further context, classifier B acheives 80% accuracy, which is higher than 75%.



Which is the “best” classifier?

Classifier 1		Class	
		Cancer	¬Cancer
Classified	Cancer	45	20
	¬Cancer	5	30

Classifier 2		Class	
		Cancer	¬Cancer
Classified	Cancer	40	10
	¬Cancer	10	40

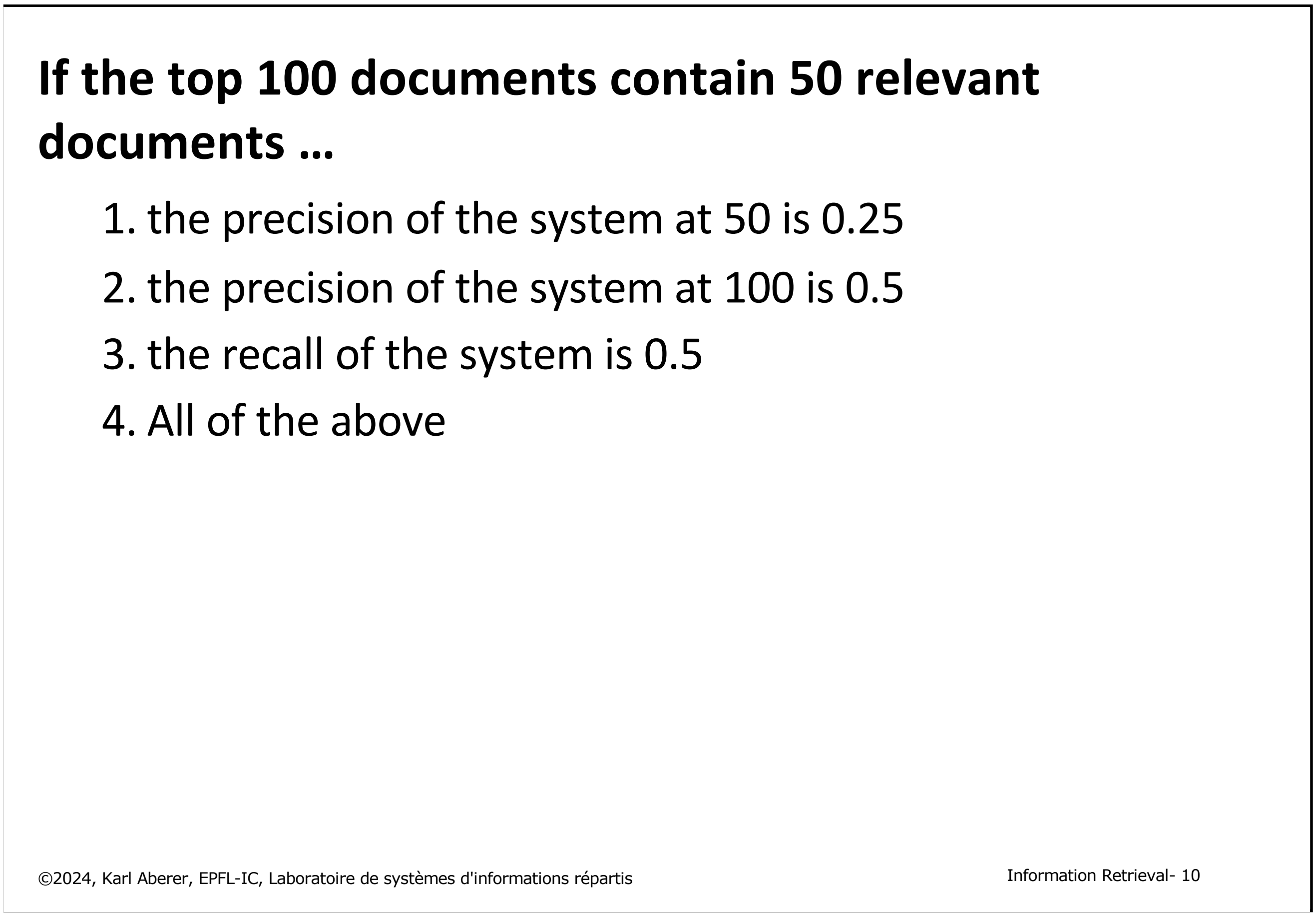
A. Classifier 1

B. Classifier 2

C. Both are equally good

Answer A

Being aware of the context, we understand that the objective is to miss as few cancers as possible. Classifier 2 misses 1 cancer out of 5, whereas cassifier 1 misses 1 out of 9, which is more suitable under this objective.

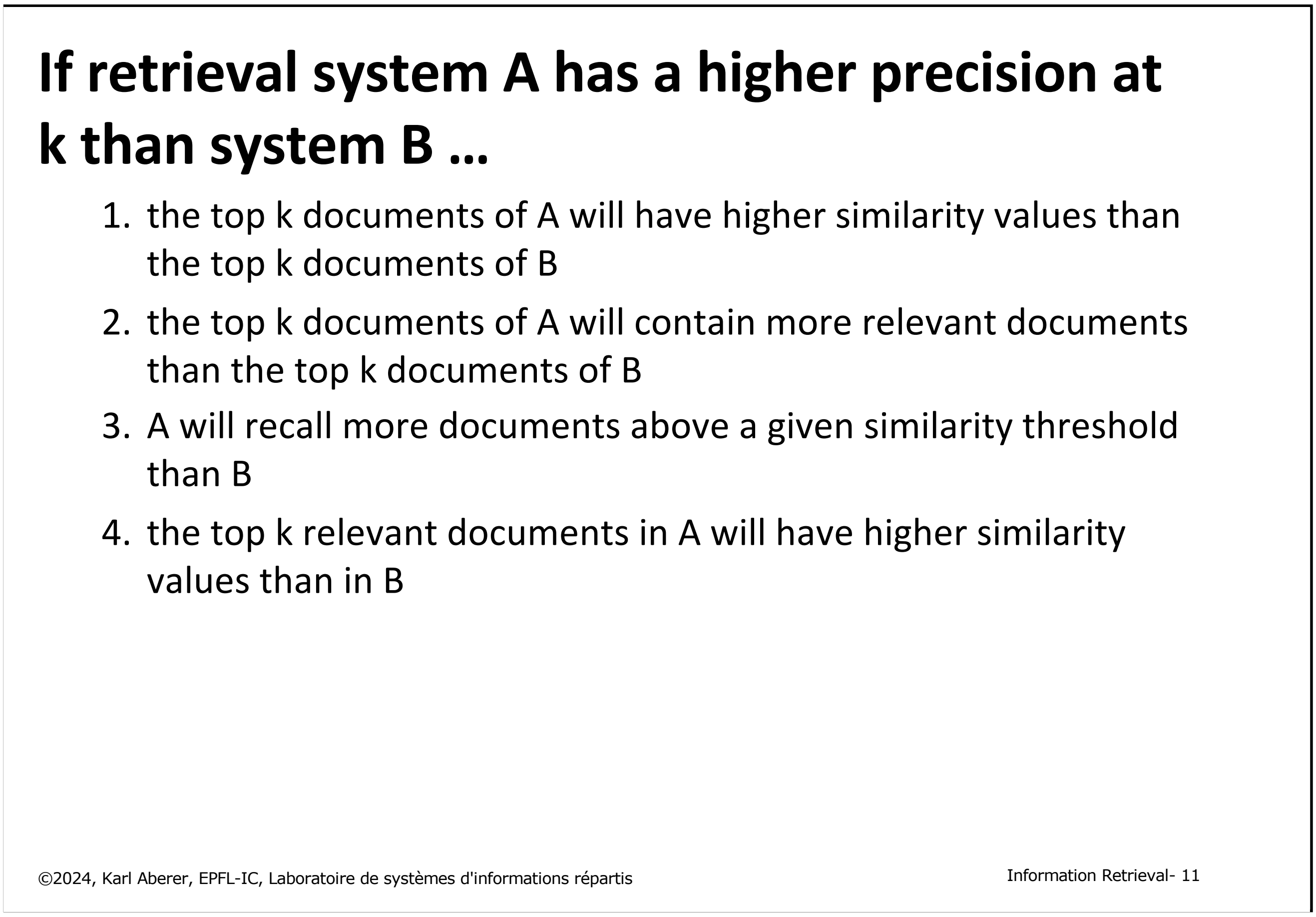


If the top 100 documents contain 50 relevant documents …

1. the precision of the system at 50 is 0.25
2. the precision of the system at 100 is 0.5
3. the recall of the system is 0.5
4. All of the above

Answer 2

Since among the retrieved 100 documents 50 are relevant, the precision at 100 is clearly 0.5. We cannot say anything about the precision at 50, since we do not know how many of the relevant docuemnts are found in the first 50 documents. We also cannot say anything about the recall, since we do not now how many relevant documents do exist.



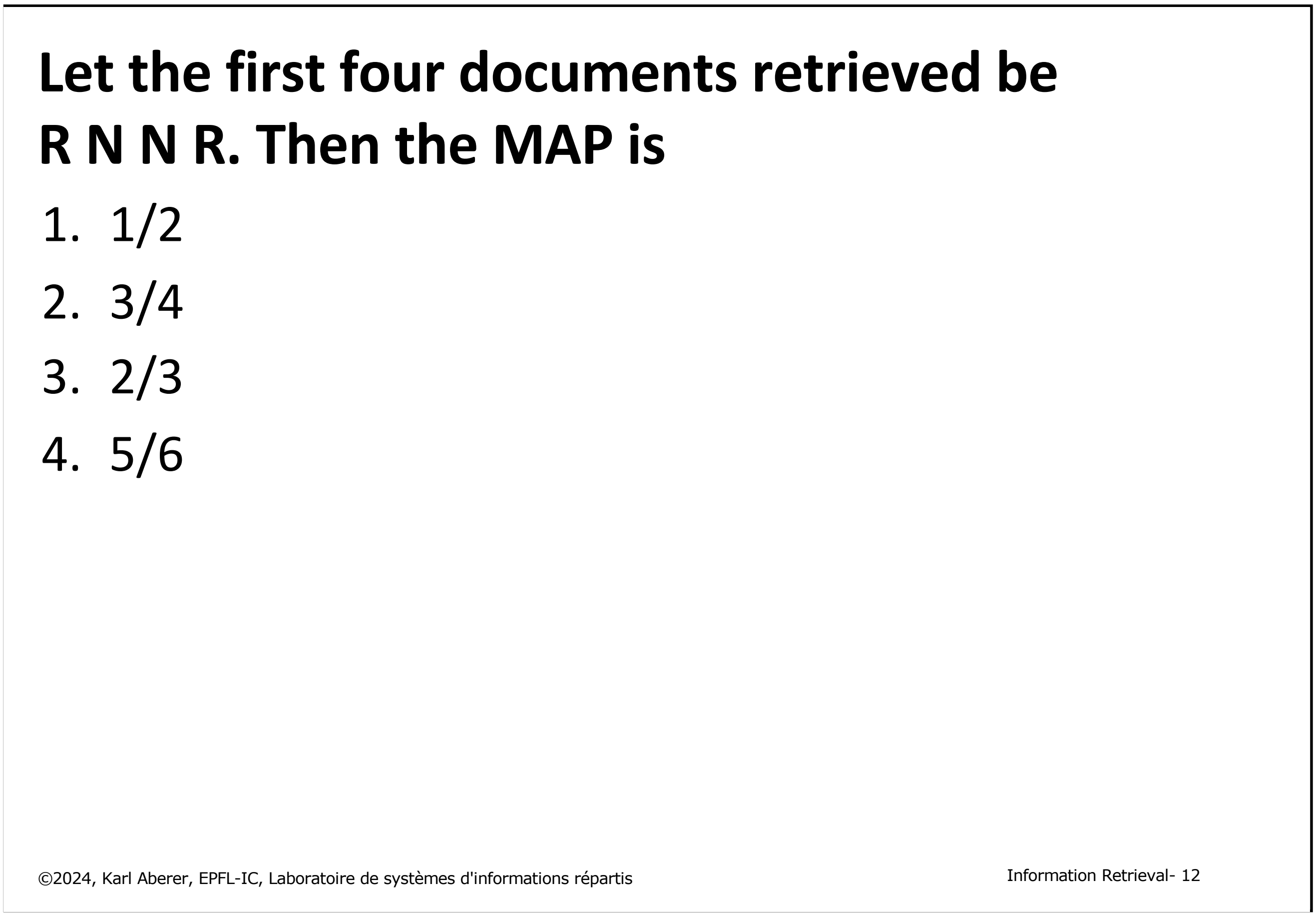
If retrieval system A has a higher precision at k than system B …

1. the top k documents of A will have higher similarity values than the top k documents of B
2. the top k documents of A will contain more relevant documents than the top k documents of B
3. A will recall more documents above a given similarity threshold than B
4. the top k relevant documents in A will have higher similarity values than in B

Information Retrieval- 11

Answer 2

If sysetm A has higher precision at k than system B, then b defition it will contain more relevant documents in the first k retrieved documents than B. The absolute similarity values that different retrieval systems produces does not give any information on how they will rank the results.



Let the first four documents retrieved be R N N R. Then the MAP is

1. 1/2
2. 3/4
3. 2/3
4. 5/6

Answer 2

P@1 = 1 and P@4 = ½. The average of the two values is this ¾.