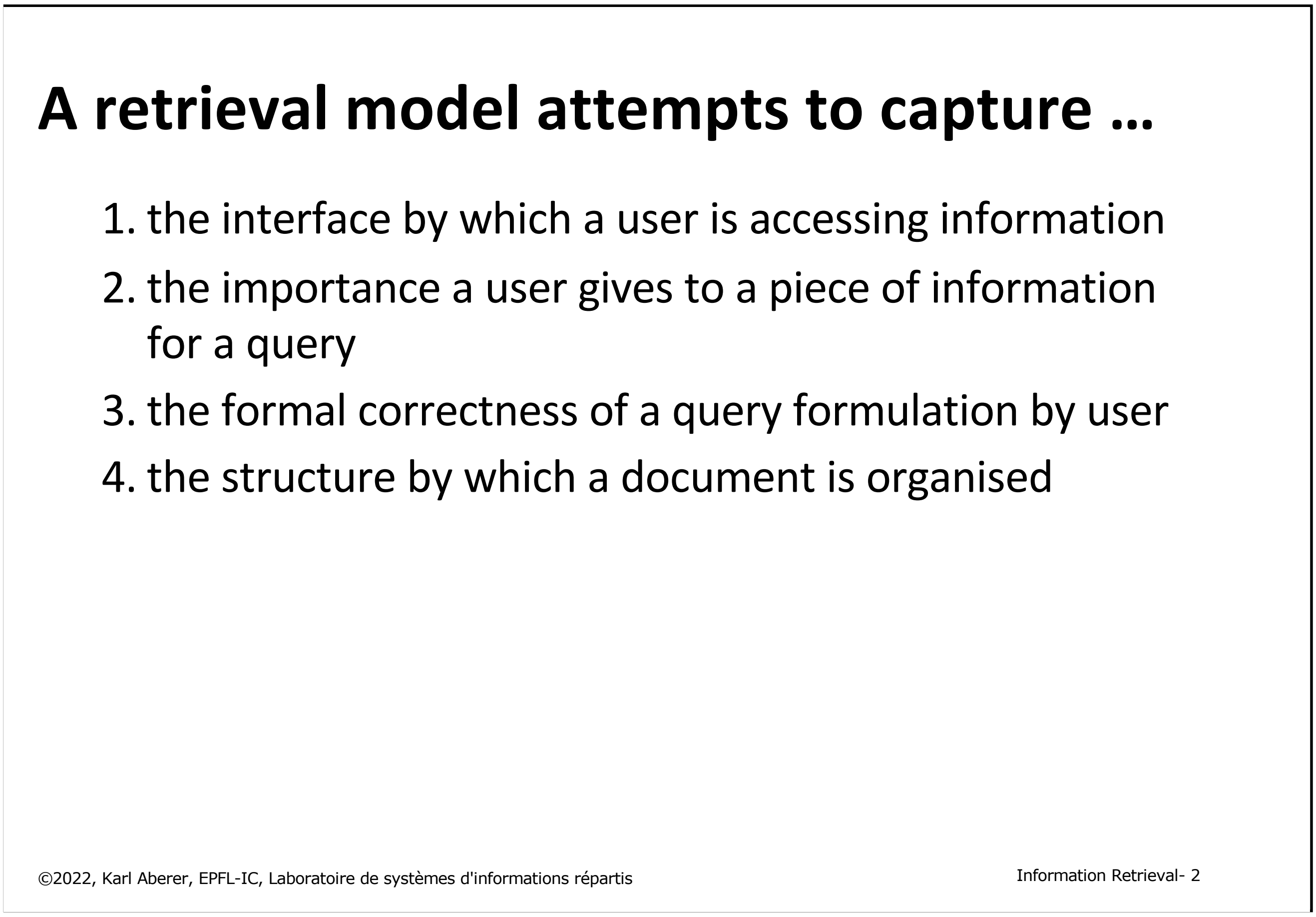
Answers to quizzes

1.1 INTRODUCTION TO INFORMATION RETRIEVAL

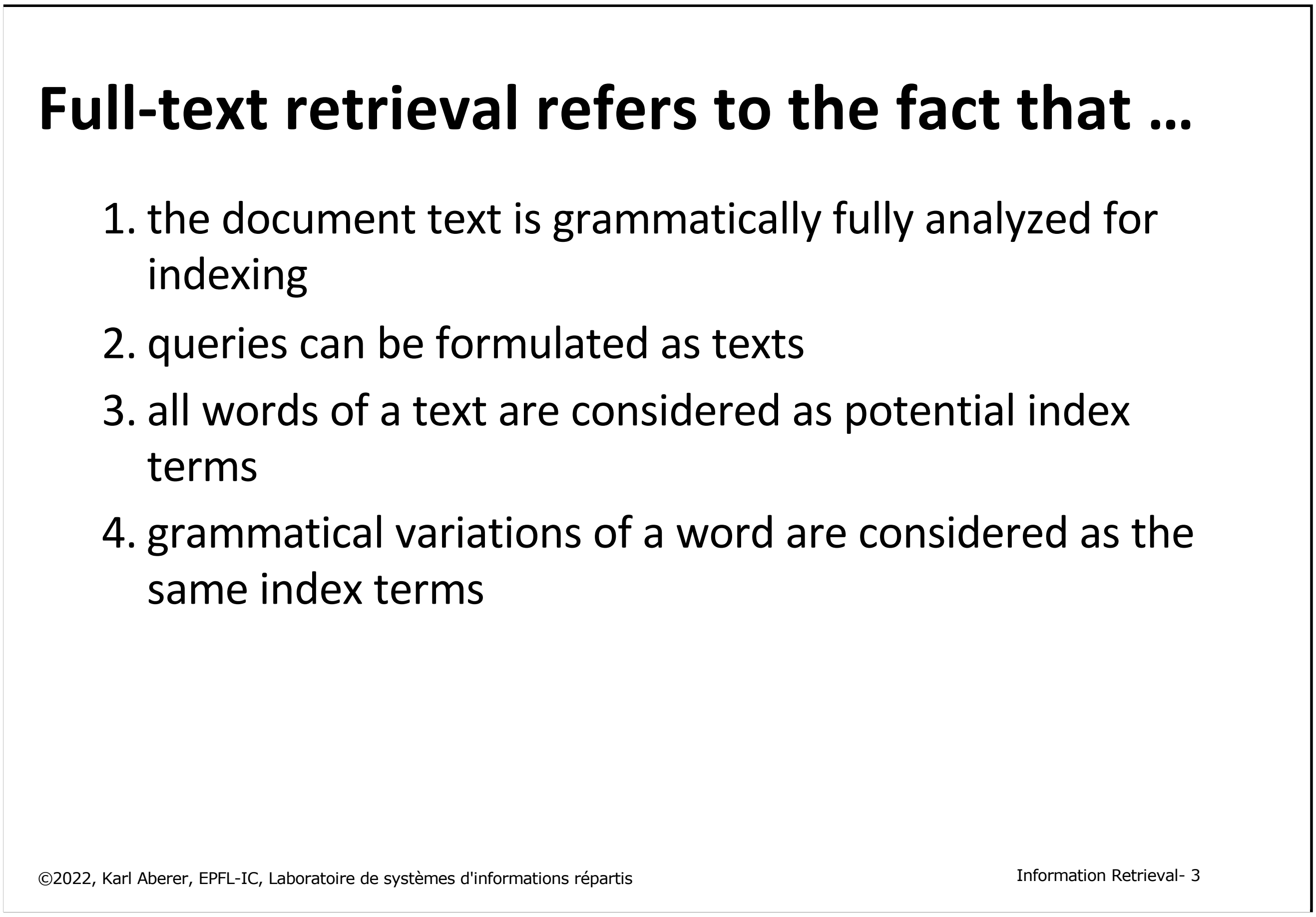


A retrieval model attempts to capture …

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.

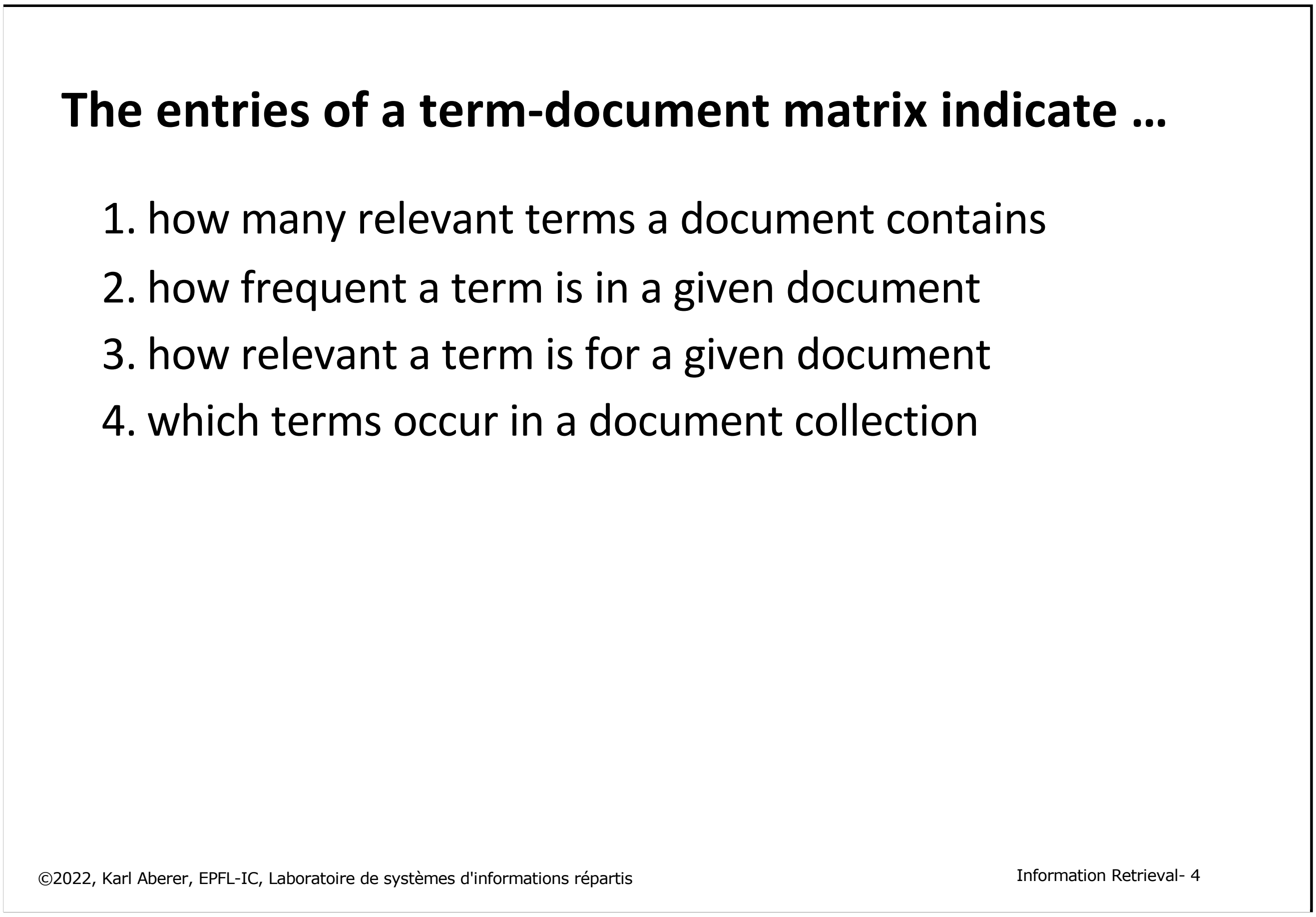


Full-text retrieval refers to the fact that …

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.

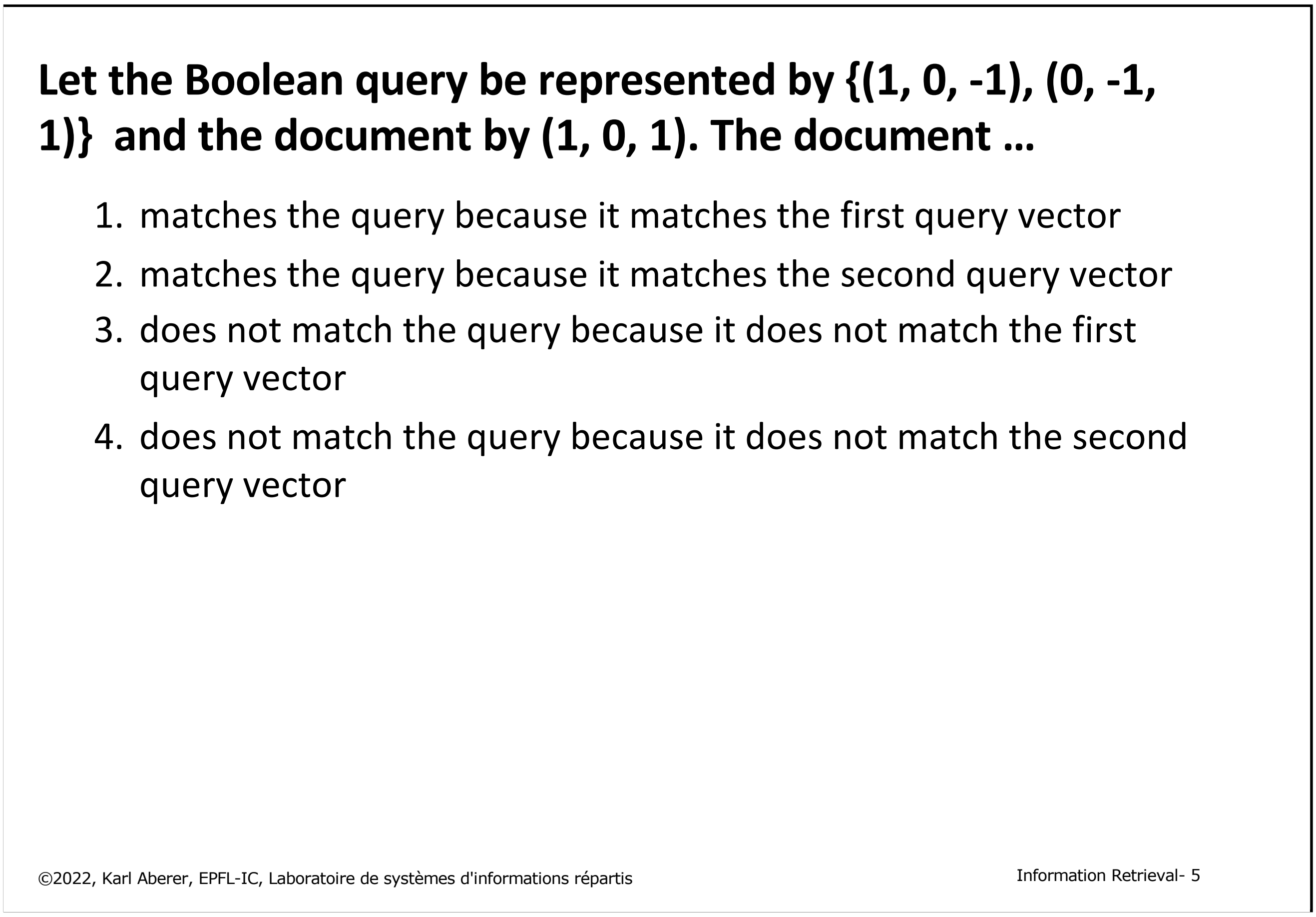


The entries of a term-document matrix indicate …

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.

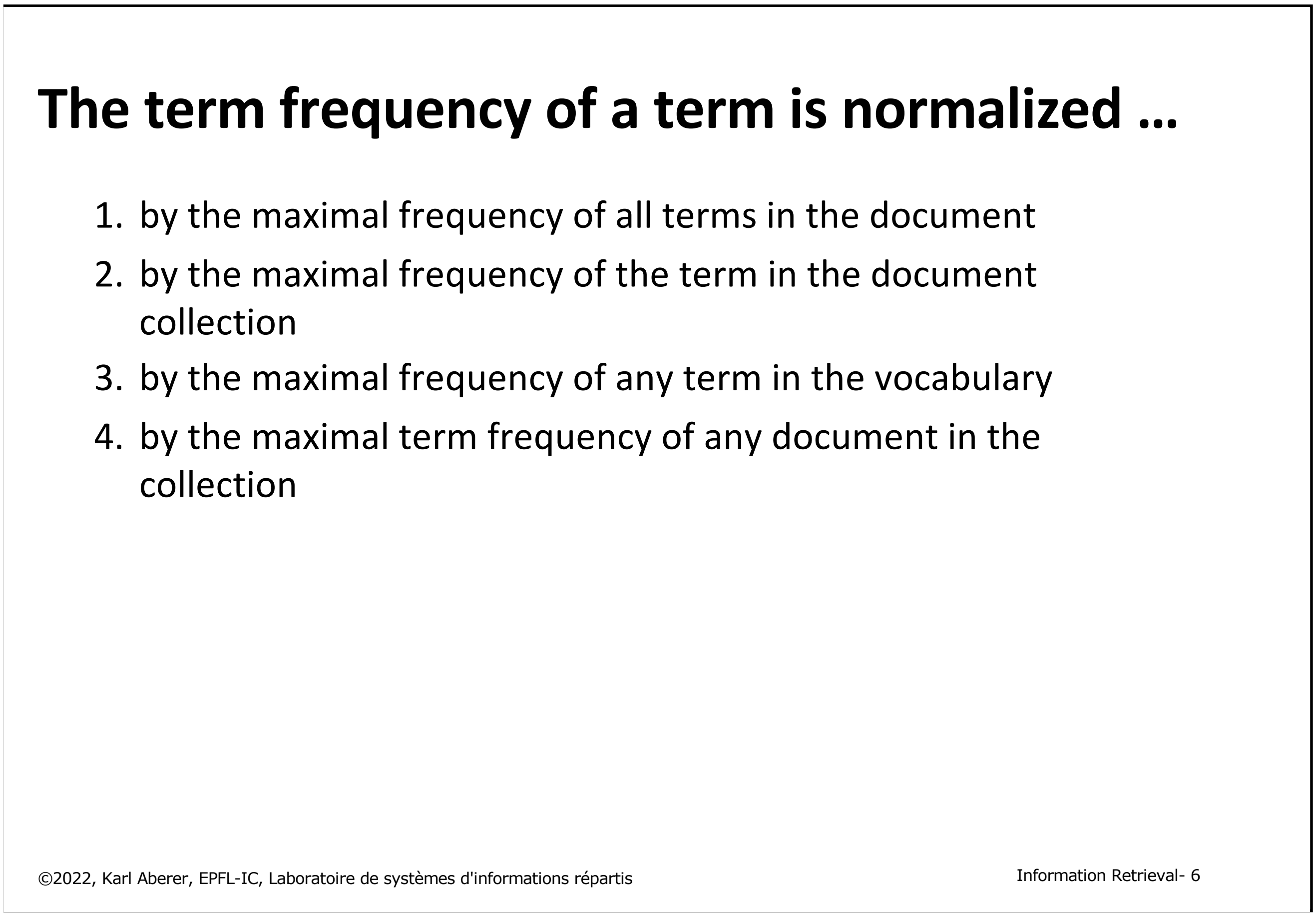


Let the Boolean query be represented by {(1, 0, -1), (0, -1, 1)} and the document by (1, 0, 1). The document …

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.

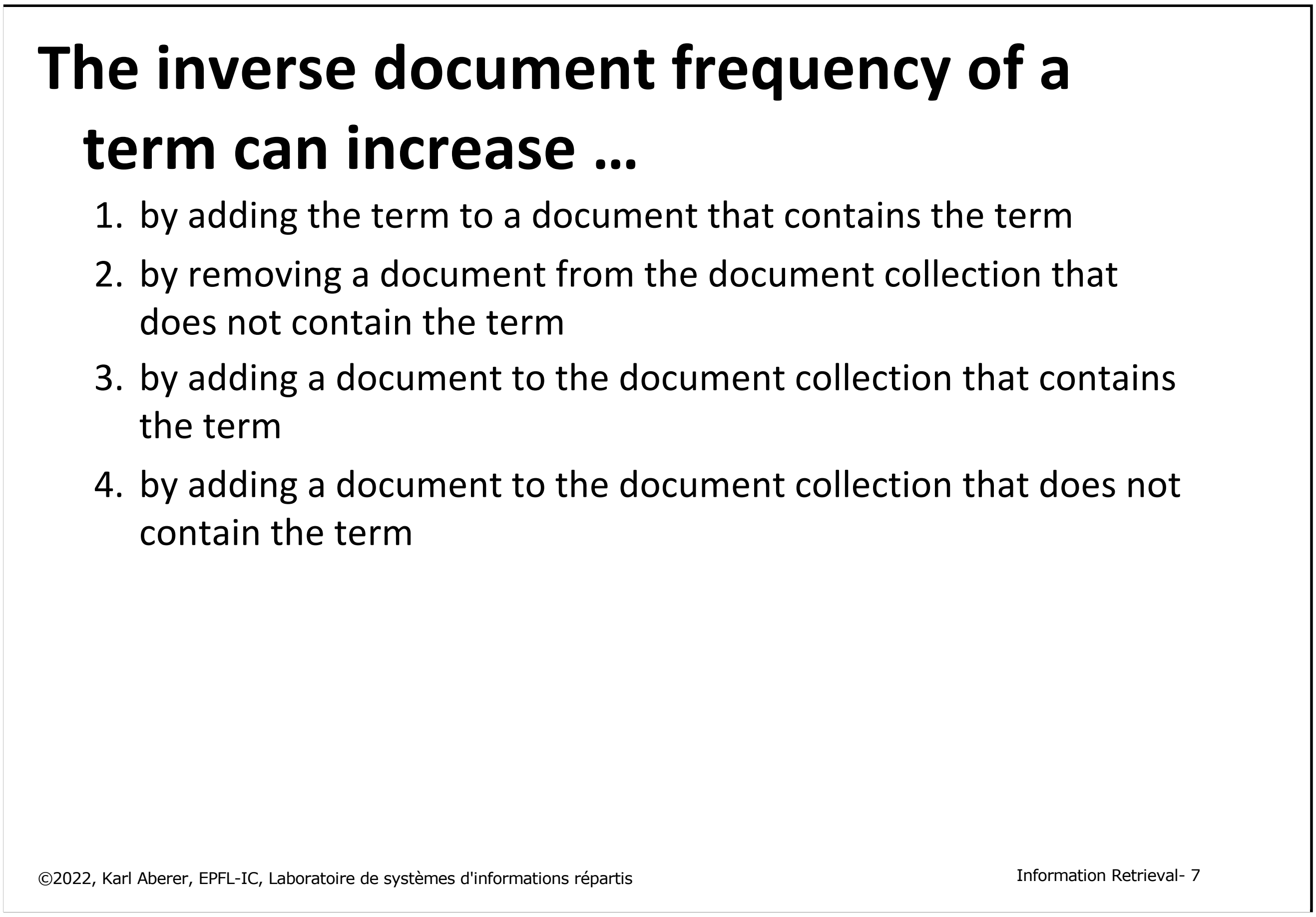


The term frequency of a term is normalized …

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 …

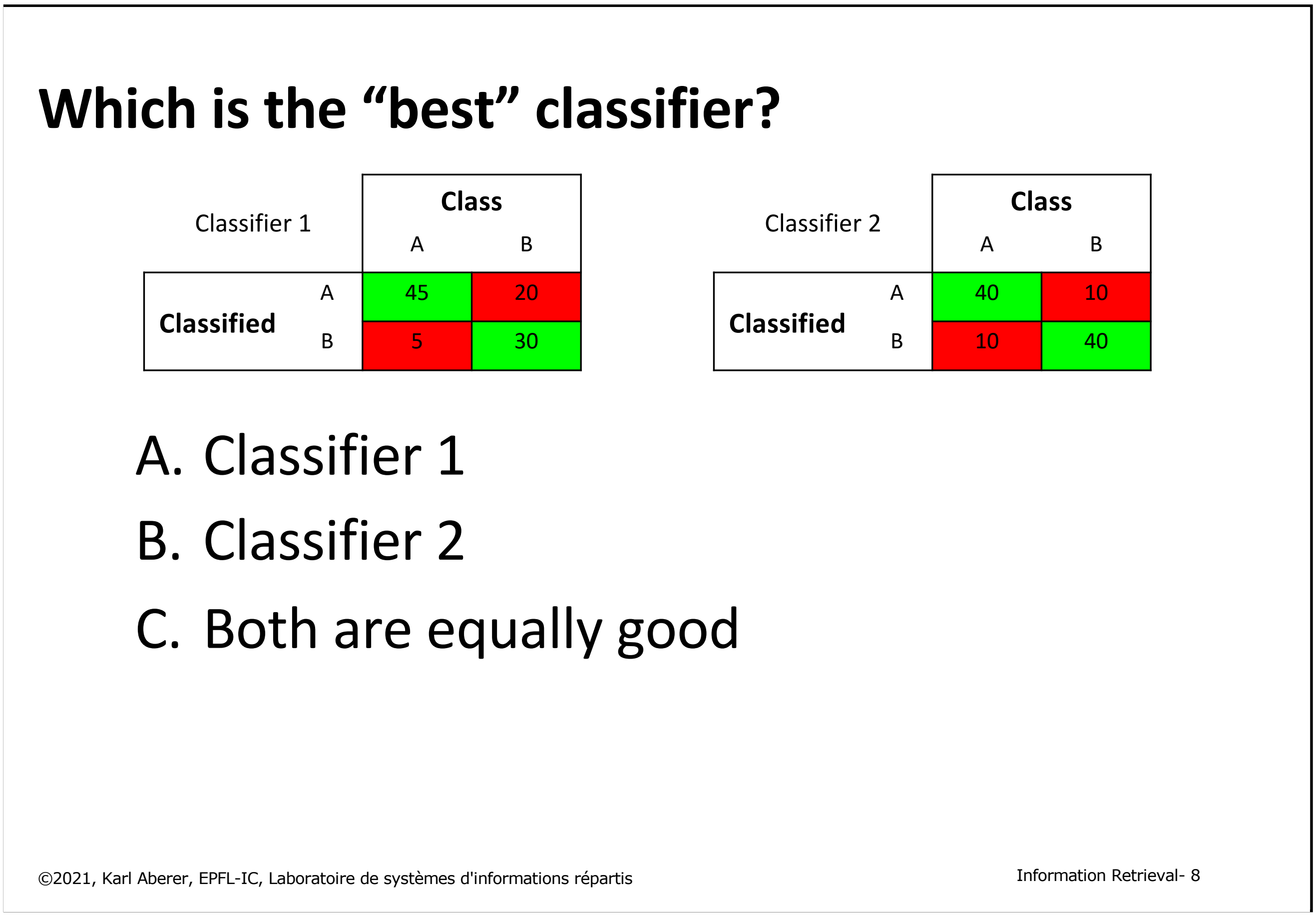
1. by adding the term to a document that contains the term
2. by removing a document from the document collection that does not contain the term
3. by adding a document to the document collection that contains the term
4. by adding a document to the document collection that does not contain the term

©2022, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Information Retrieval- 7

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

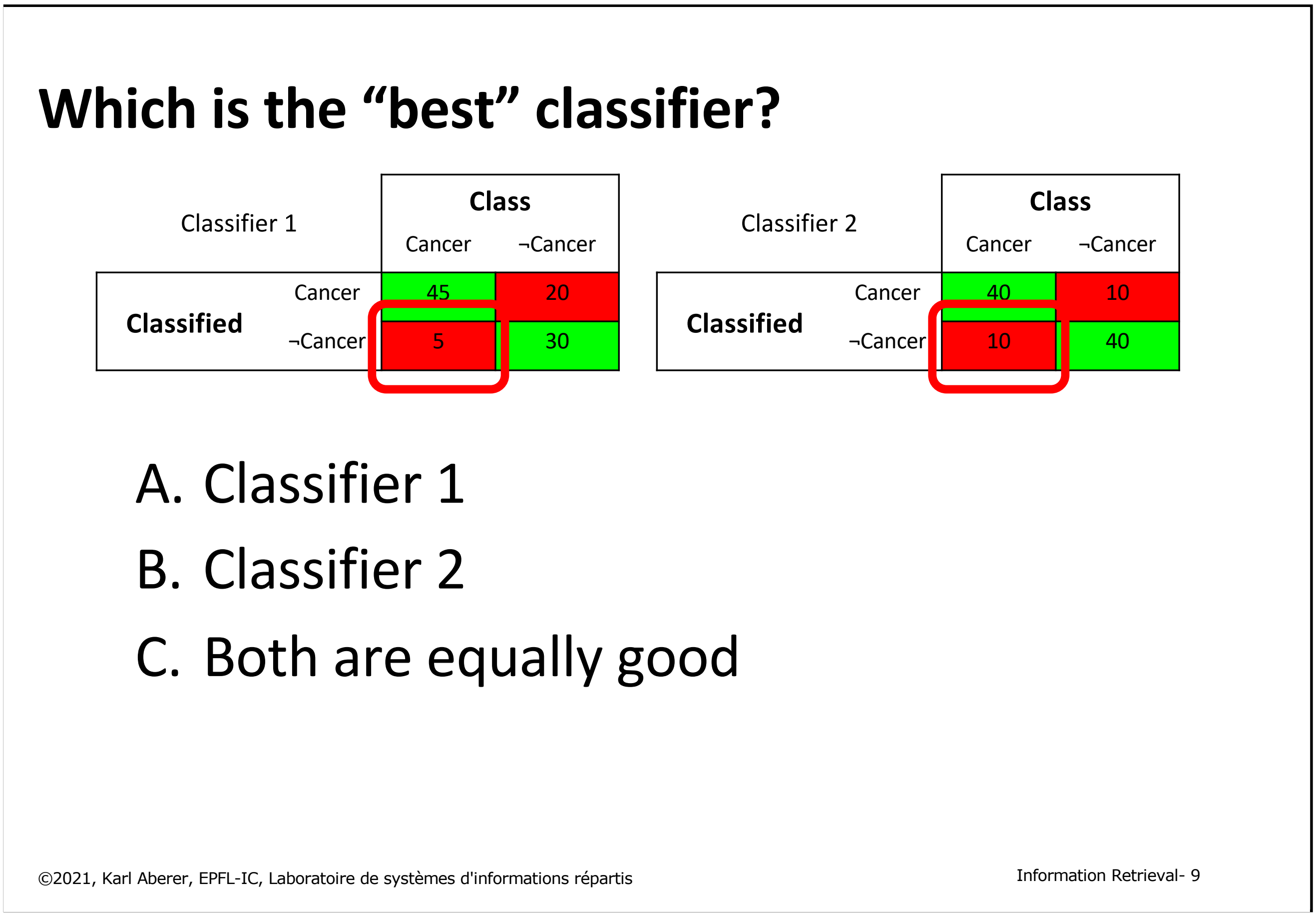
A. Classifier 1

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

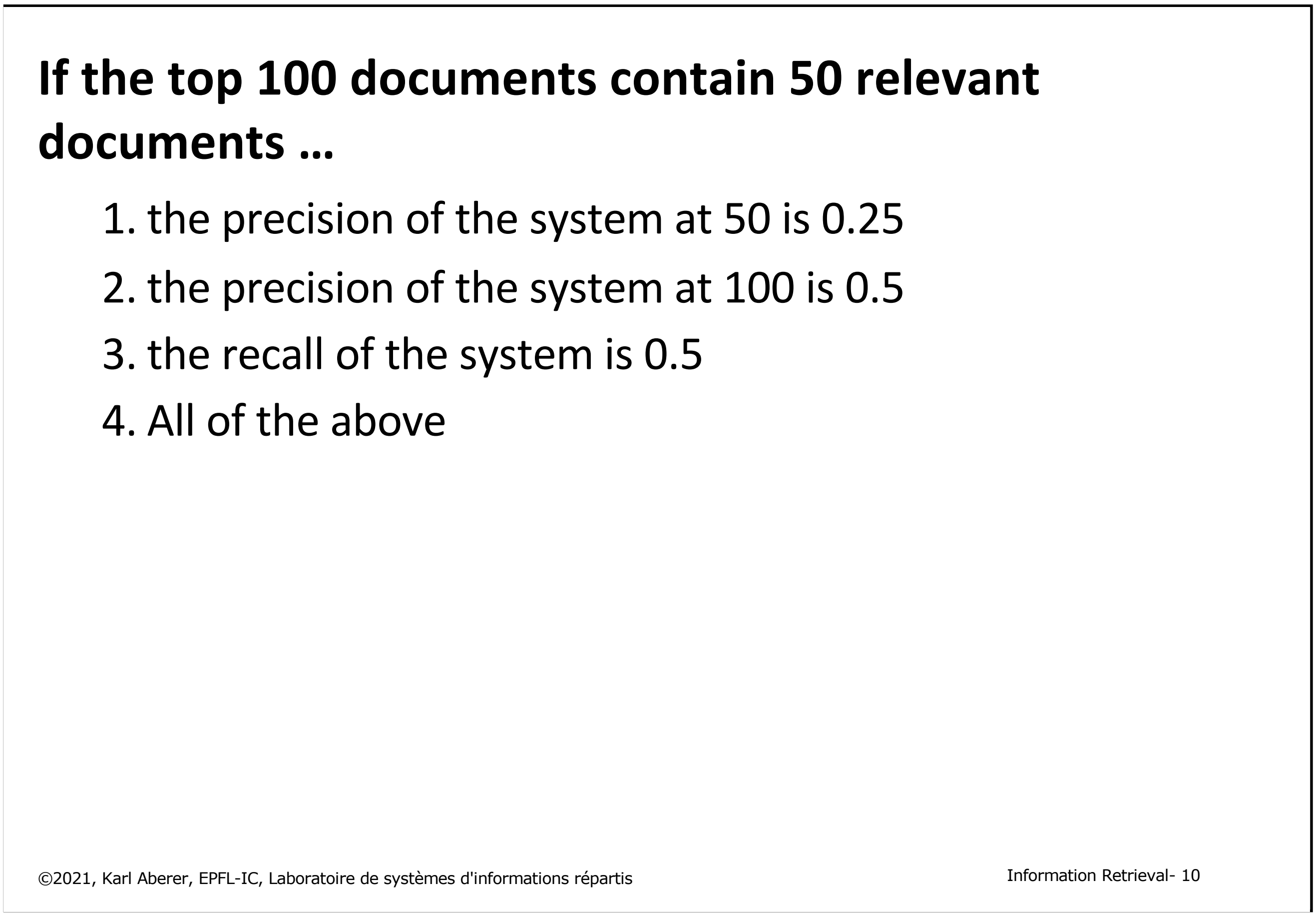
A. Classifier 1

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.

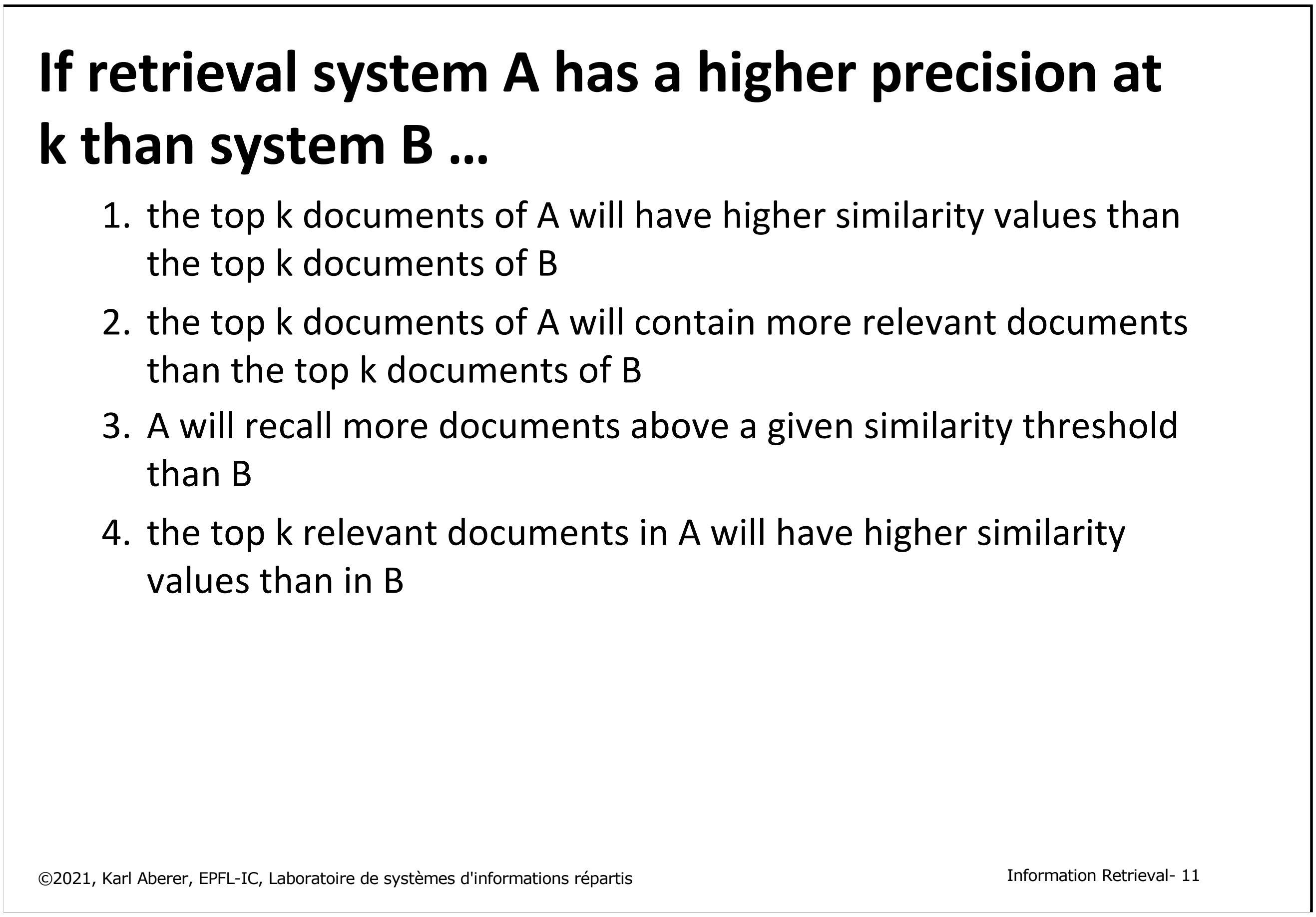


If the top 100 documents contain 50 relevant documents …

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.

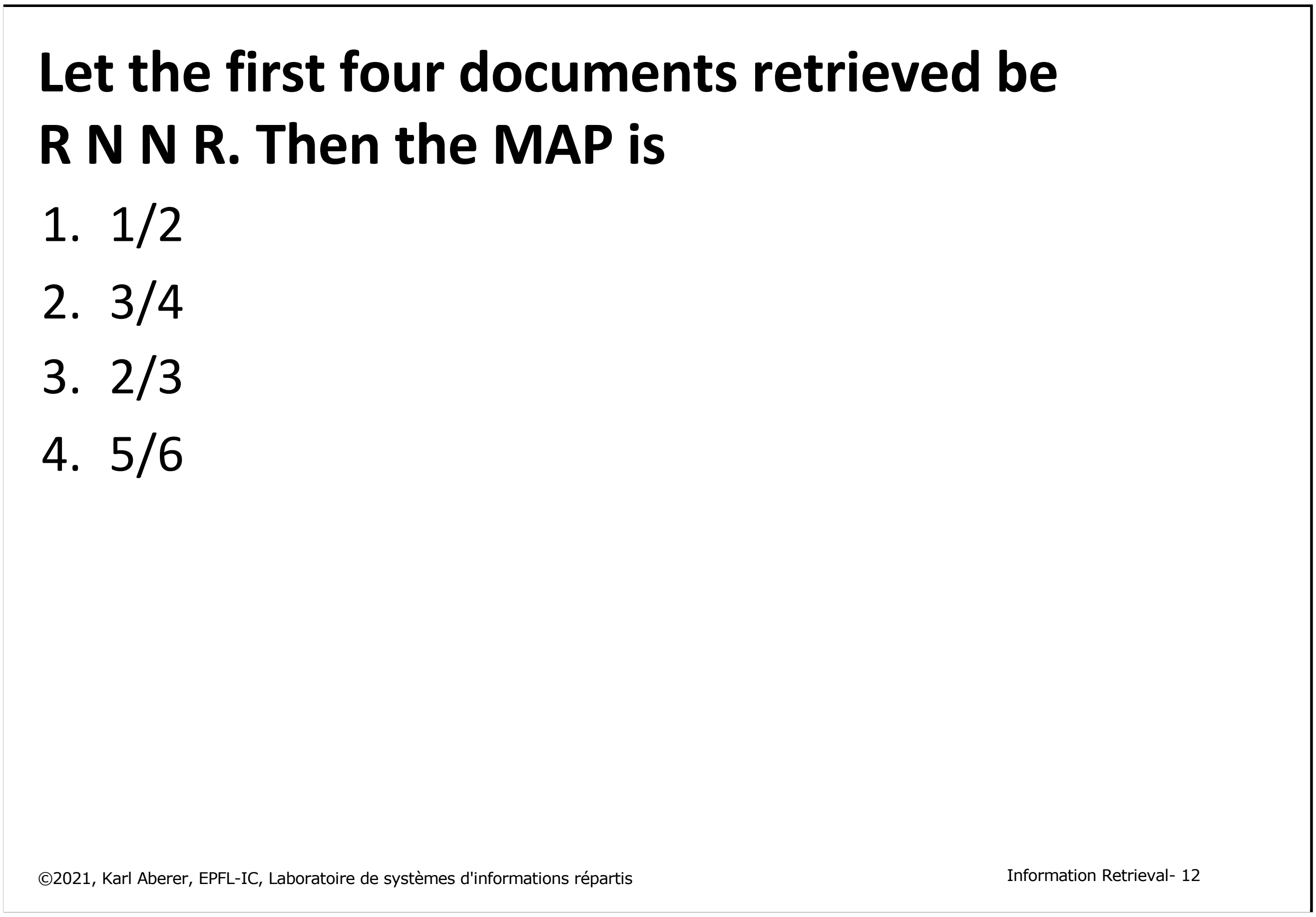


If retrieval system A has a higher precision at k than system B …

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

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.

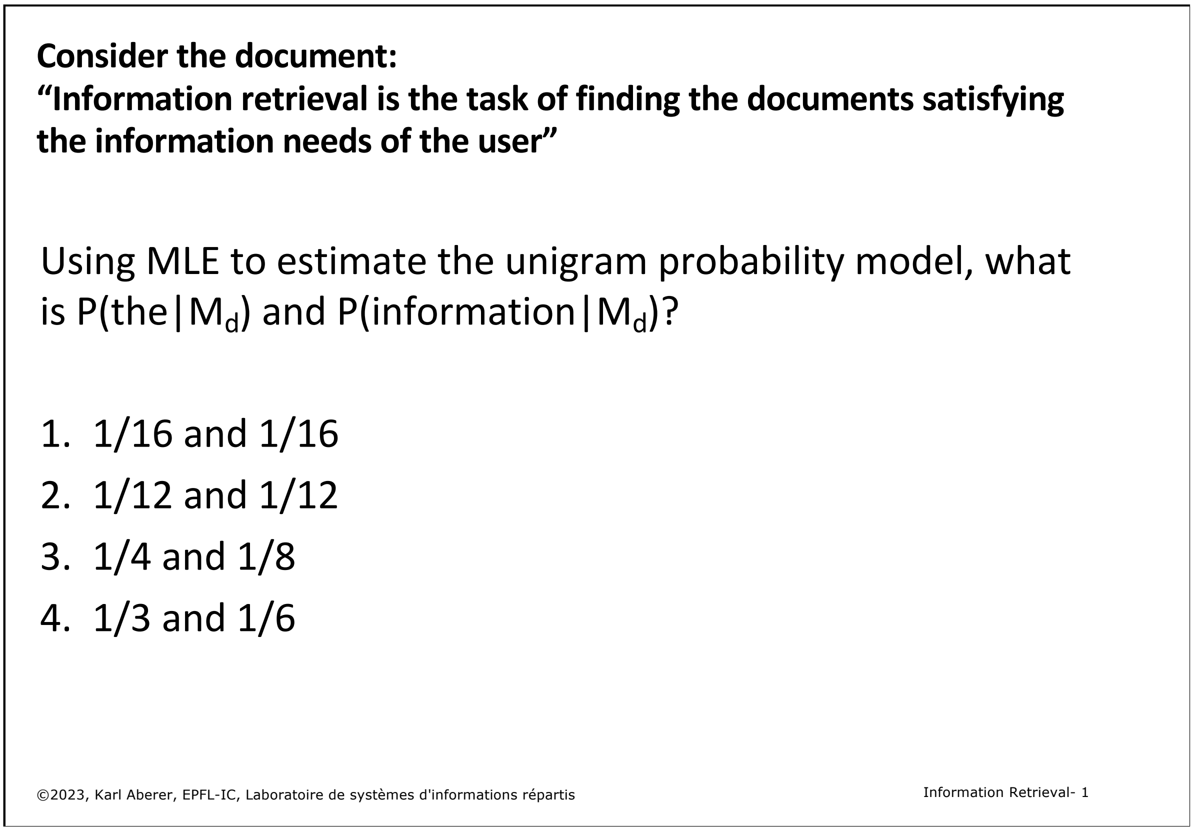


Let the first four documents retrieved be R N N R. Then the MAP is

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

Anwer 2

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



Consider the document:
“Information retrieval is the task of finding the documents satisfying the information needs of the user”

Using MLE to estimate the unigram probability model, what is P(the|M_d) and P(information|M_d)?

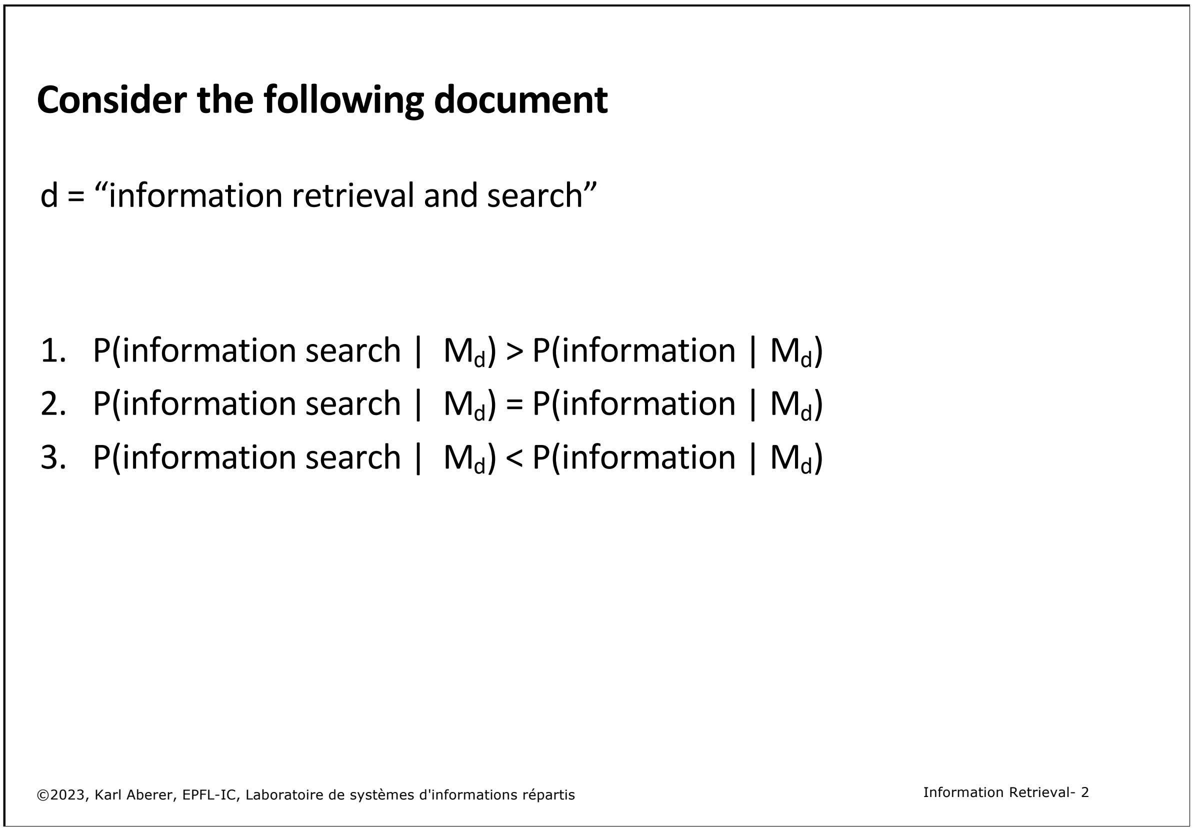
1. 1/16 and 1/16
2. 1/12 and 1/12
3. 1/4 and 1/8
4. 1/3 and 1/6

Answer 3

The total number of terms in the document is 16.

The term ”the” appears 4 times, the term information 2 times.

Therefore the probabilities are 4/16 and 2/16.



Consider the following document

d = “information retrieval and search”

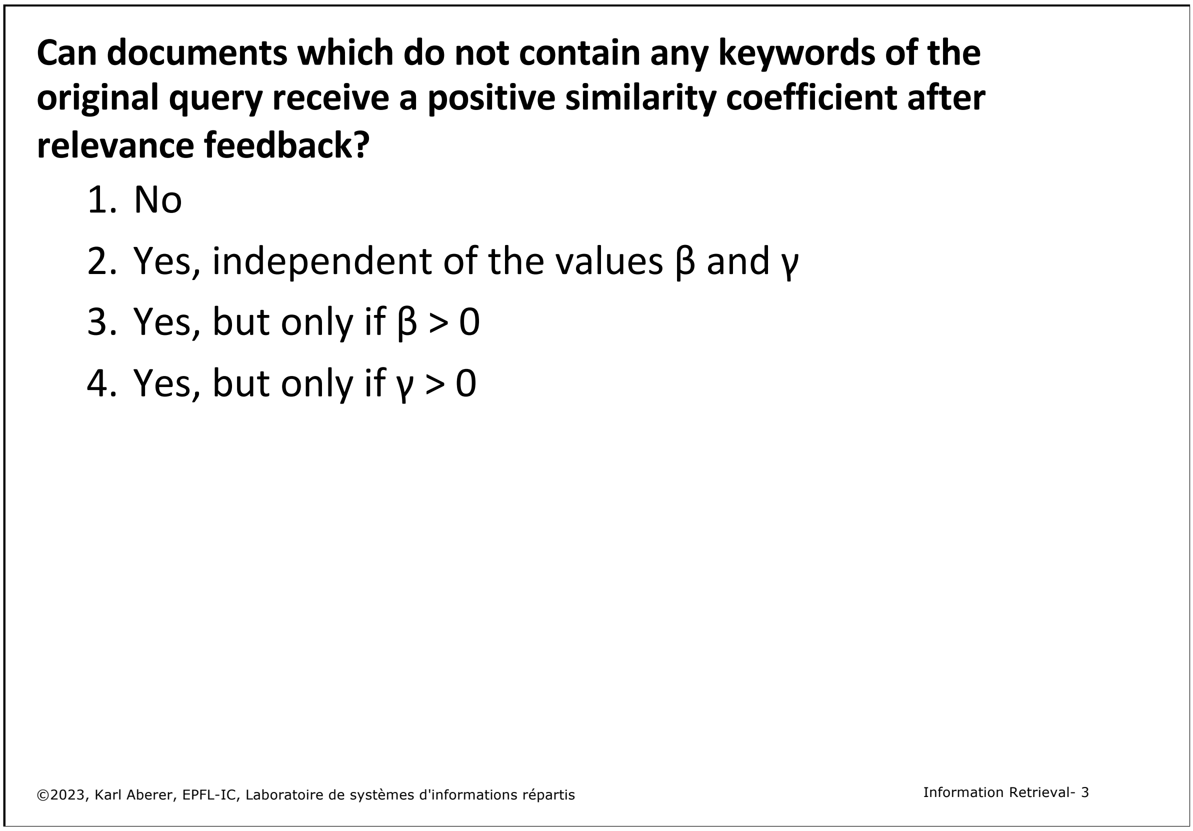
1. P(information search | M_d) > P(information | M_d)
2. P(information search | M_d) = P(information | M_d)
3. P(information search | M_d) < P(information | M_d)

Answer 3

The probability P(t1 t2 | M_d) is the product of the probabilities P(t1 | M) and P(t2 | M).

Since P(t2 | M) < 1, necessarily P(t1 t2 | M_d) < P(t1| M_d)

Note that this is not a problem in the sense that we are never comparing probabilities of different queries in retrieval, but probabilities of different documents for the same query.

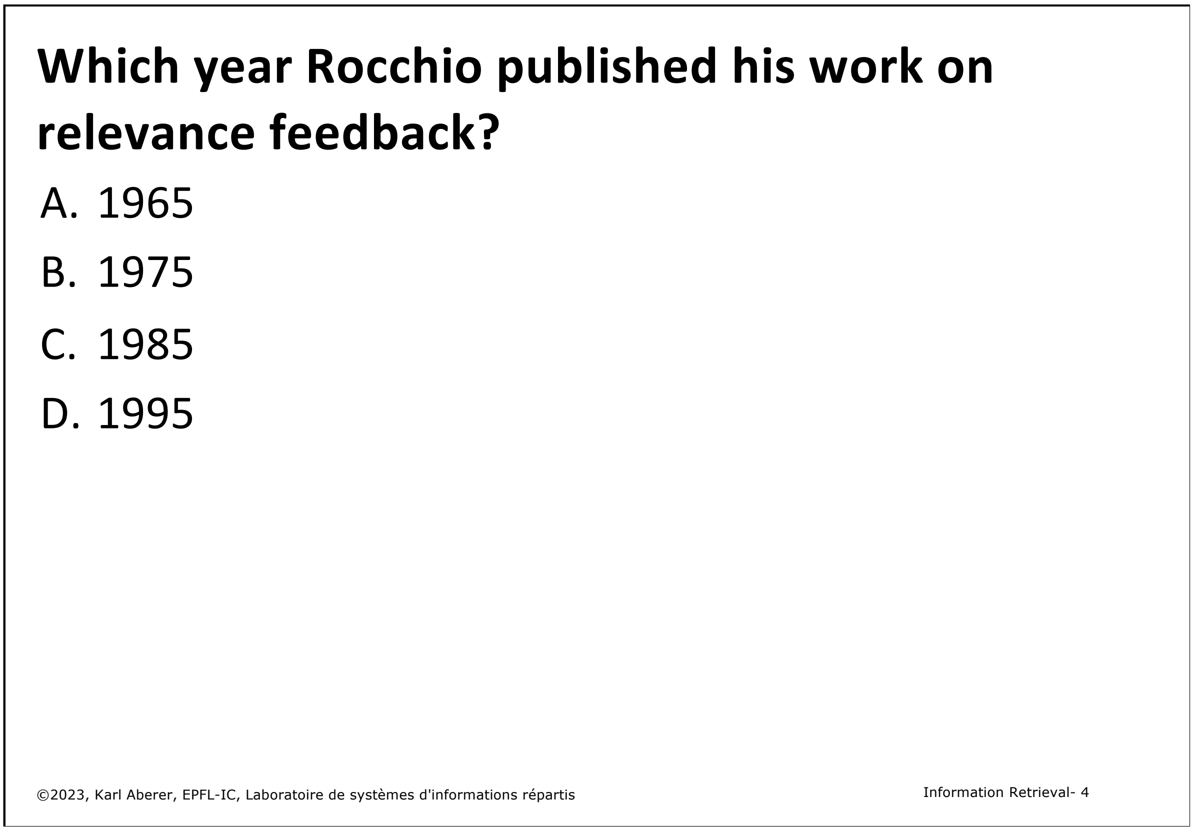


Can documents which do not contain any keywords of the original query receive a positive similarity coefficient after relevance feedback?

1. No
2. Yes, independent of the values β and γ
3. Yes, but only if β > 0
4. Yes, but only if γ > 0

Answer 3

It is possible that a document that does not contain any terms of the original query can contain terms of another document that has been selected for the set of relevant documents D_r. If β > 0. This will result in a positive similarity value. Note that all weights in document and query vectors are positive, so the scalar product of two such vectors will be positive.



Which year Rocchio published his work on relevance feedback?

A. 1965

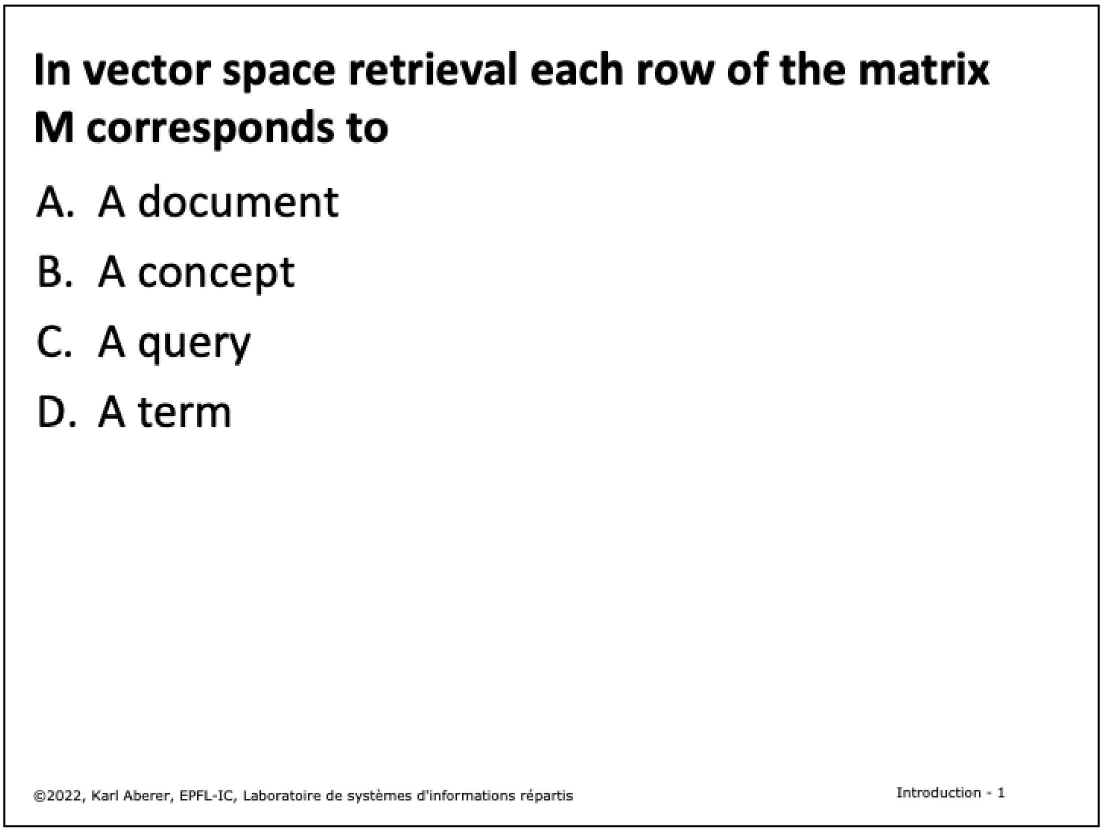
B. 1975

C. 1985

D. 1995

Answer A

Rocchio, J. J., and Gerard Salton. "Information search optimization and interactive retrieval techniques." *Proceedings of the November 30--December 1, 1965, fall joint computer conference, part I*. 1965.



In vector space retrieval each row of the matrix M corresponds to

A. A document

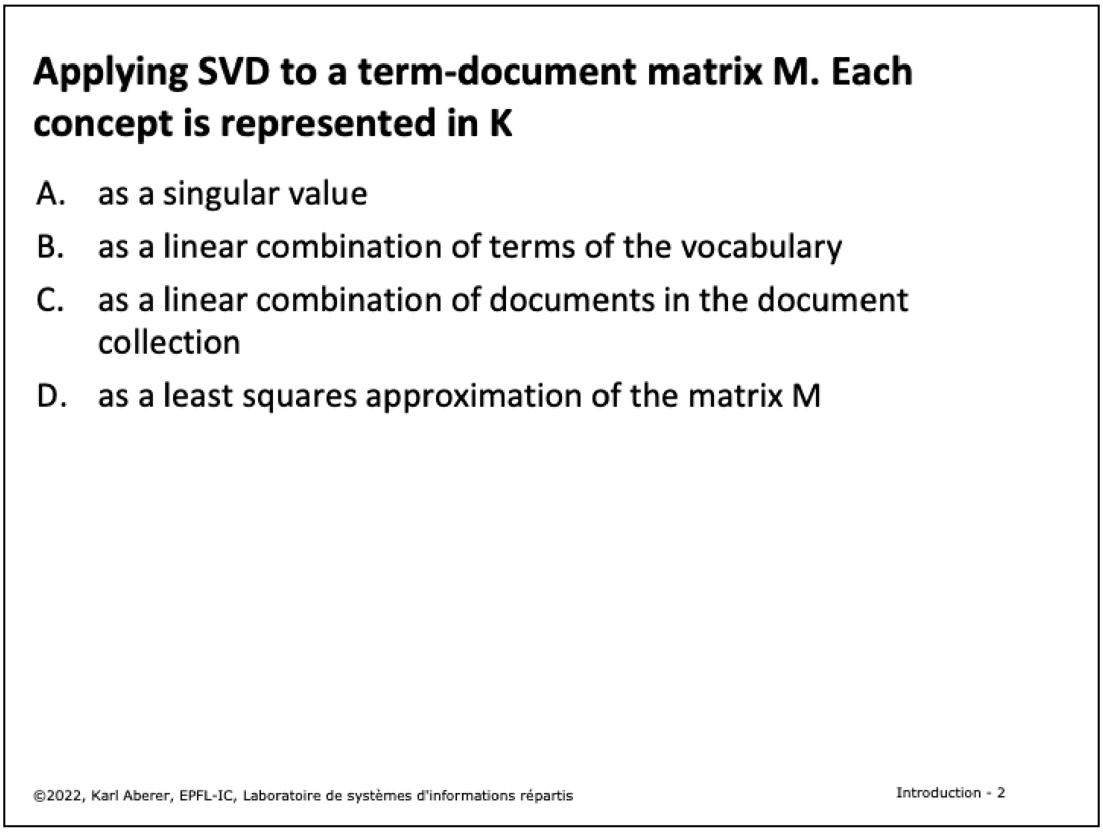
B. A concept

C. A query

D. A term

Answer D

M is an mxn matrix, with m rows corresponding to the m terms in the vocabulary.



Applying SVD to a term-document matrix M. Each concept is represented in K

A. as a singular value

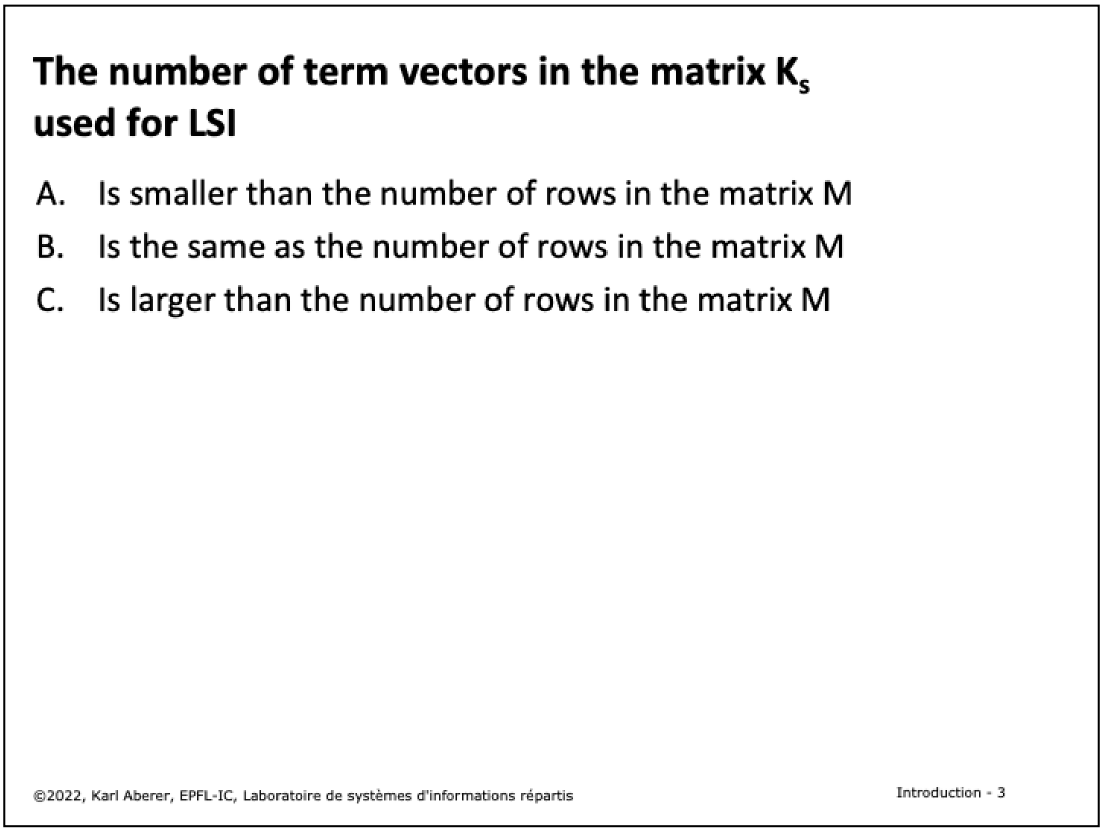
B. as a linear combination of terms of the vocabulary

C. as a linear combination of documents in the document collection

D. as a least squares approximation of the matrix M

Answer B

K is mxr matrix, where the columns correspond to vectors. These vectors correspond to a linear combination of the m terms of the vocabulary.



The number of term vectors in the matrix K_s used for LSI

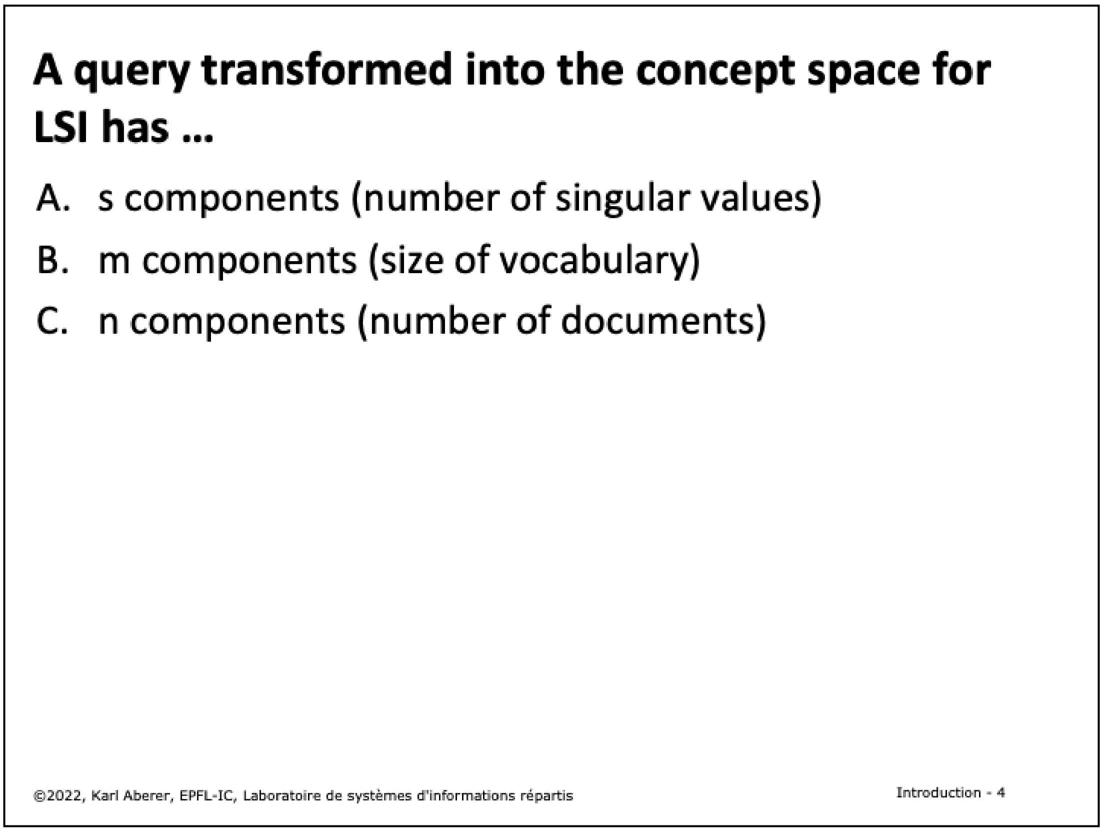
A. Is smaller than the number of rows in the matrix M

B. Is the same as the number of rows in the matrix M

C. Is larger than the number of rows in the matrix M

Answer B

K_s is a mxs matrix, where each row corresponds to a term in the vocabulary, as for M. The number of columns s is smaller than the number of columns in the original matrix K.



A query transformed into the concept space for LSI has …

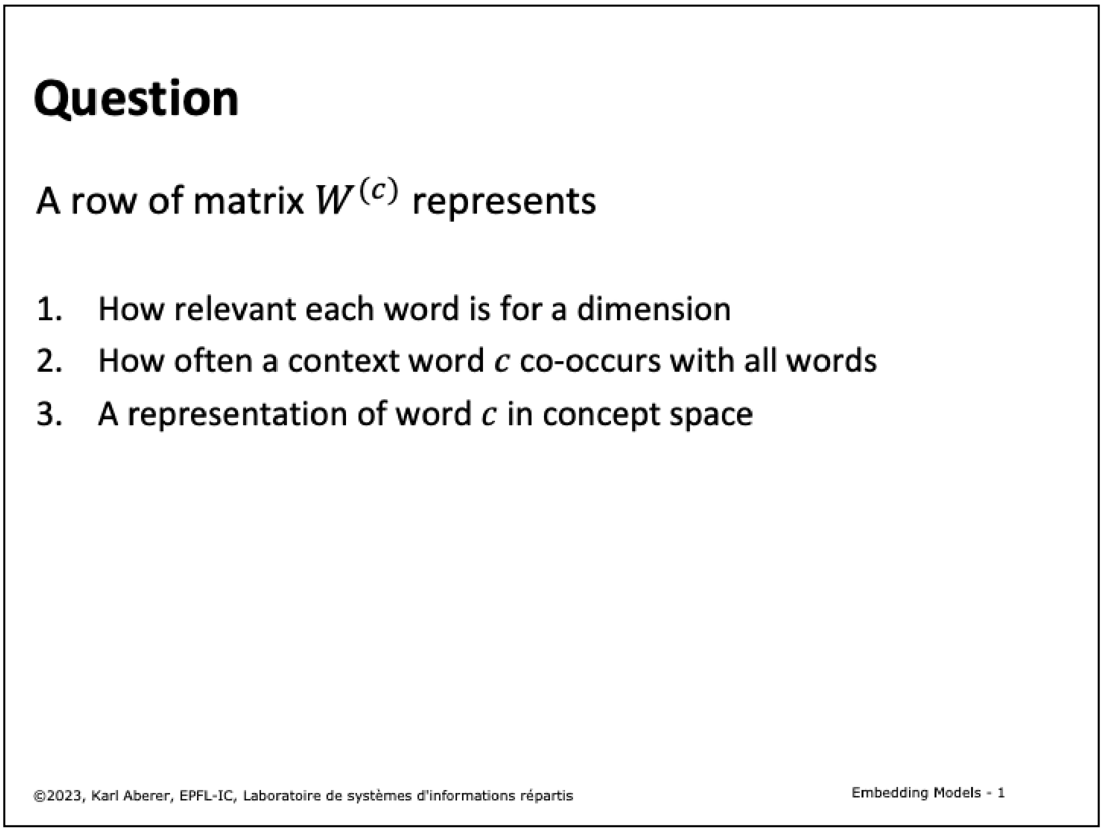
A. s components (number of singular values)

B. m components (size of vocabulary)

C. n components (number of documents)

Answer A

The transformed query is a vector over the number of selected concepts s.



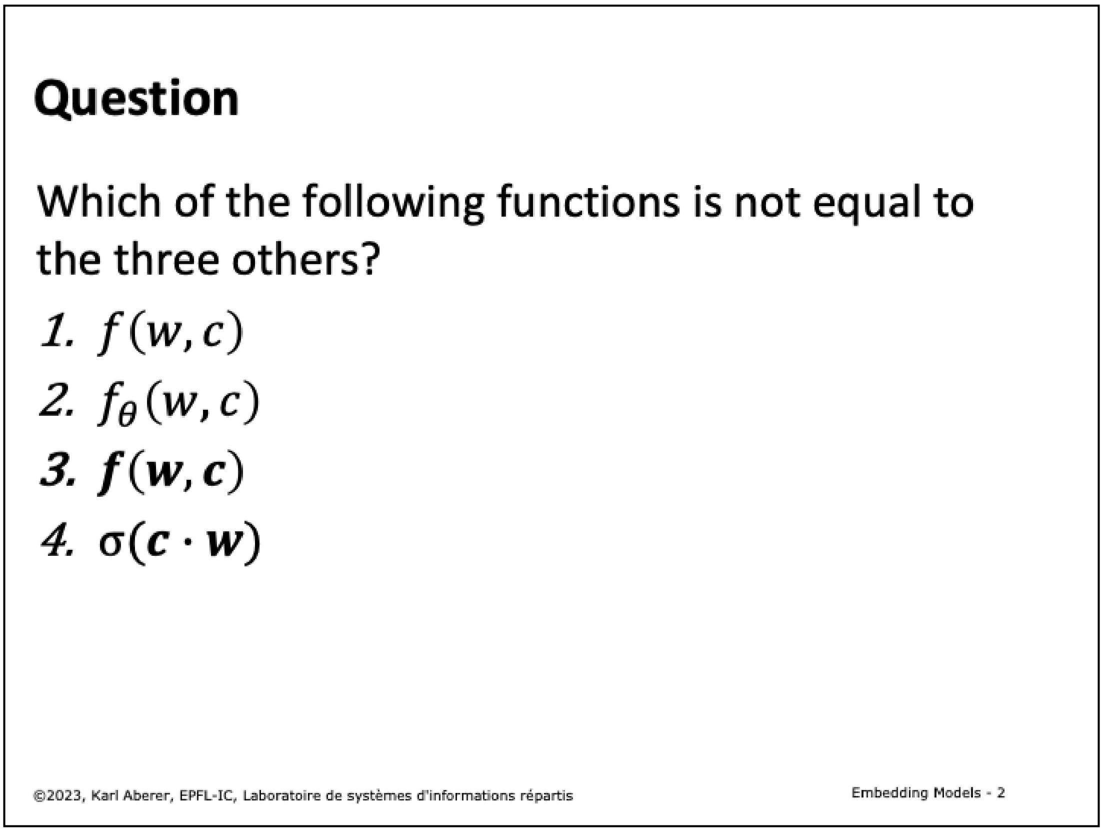
Question

A row of matrix $W^{(c)}$ represents

1. How relevant each word is for a dimension
2. How often a context word c co-occurs with all words
3. A representation of word c in concept space

Answer 1

The rows of the matrix correspond to the dimensions of the embedding. Each entry in the row corresponds to a word from the vocabulary. Therefore the row represents the importance of each word for a given dimension.



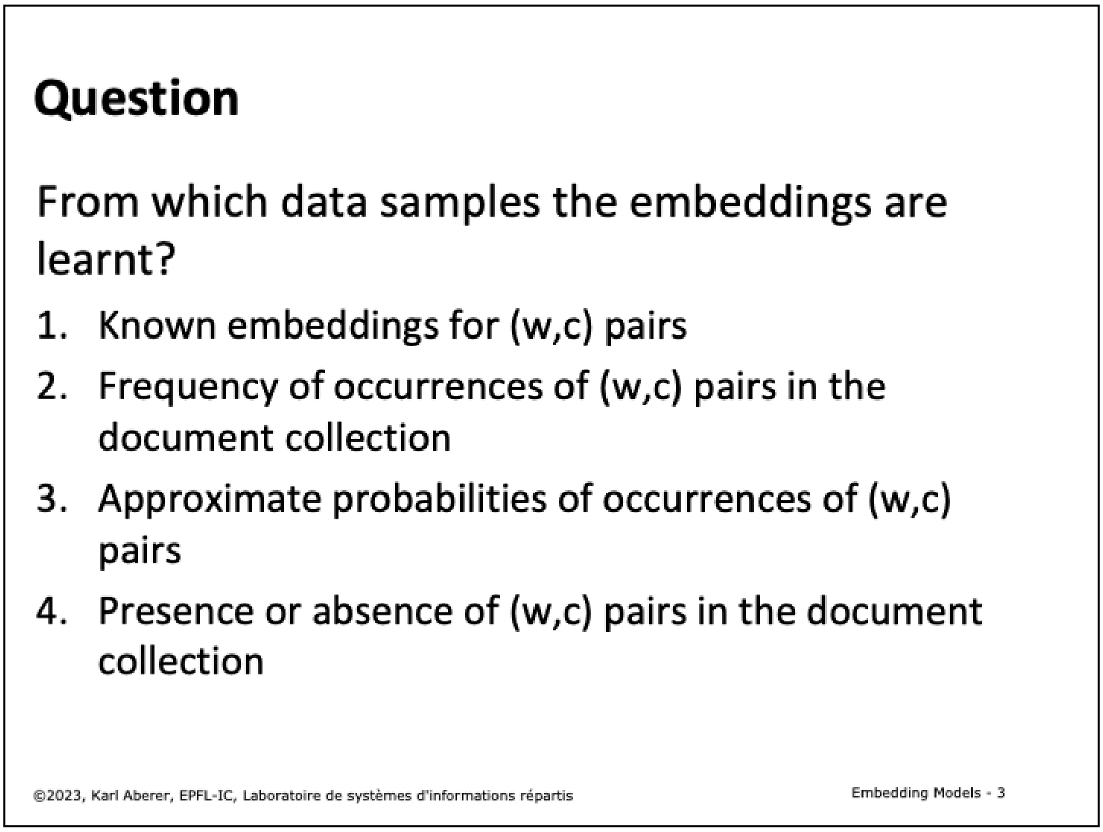
Question

Which of the following functions is not equal to the three others?

1. $f(w,c)$
2. $f_\theta(w,c)$
3. $\boldsymbol{f}(\boldsymbol{w},\boldsymbol{c})$
4. $\sigma(\boldsymbol{c}\cdot\boldsymbol{w})$

Answer 1

$f(w,c)$ is the function to be approximated. The three other functions are the same, with a more detailed specification for answers 2, 3 and 4.



Question

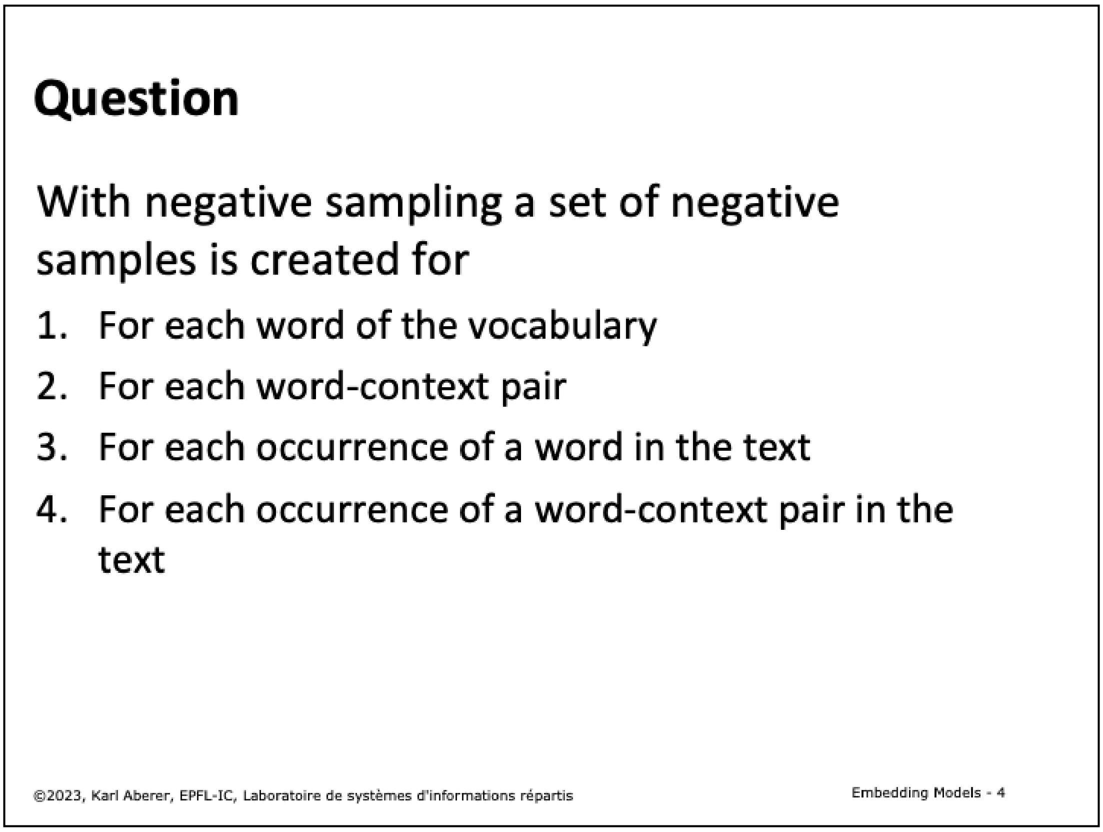
From which data samples the embeddings are learnt?

1. Known embeddings for (w,c) pairs
2. Frequency of occurrences of (w,c) pairs in the document collection
3. Approximate probabilities of occurrences of (w,c) pairs
4. Presence or absence of (w,c) pairs in the document collection

Embedding Models - 3

Answer 4

In the skipgram model the sample data consists of word-context pairs that are present or absent in the document collection.



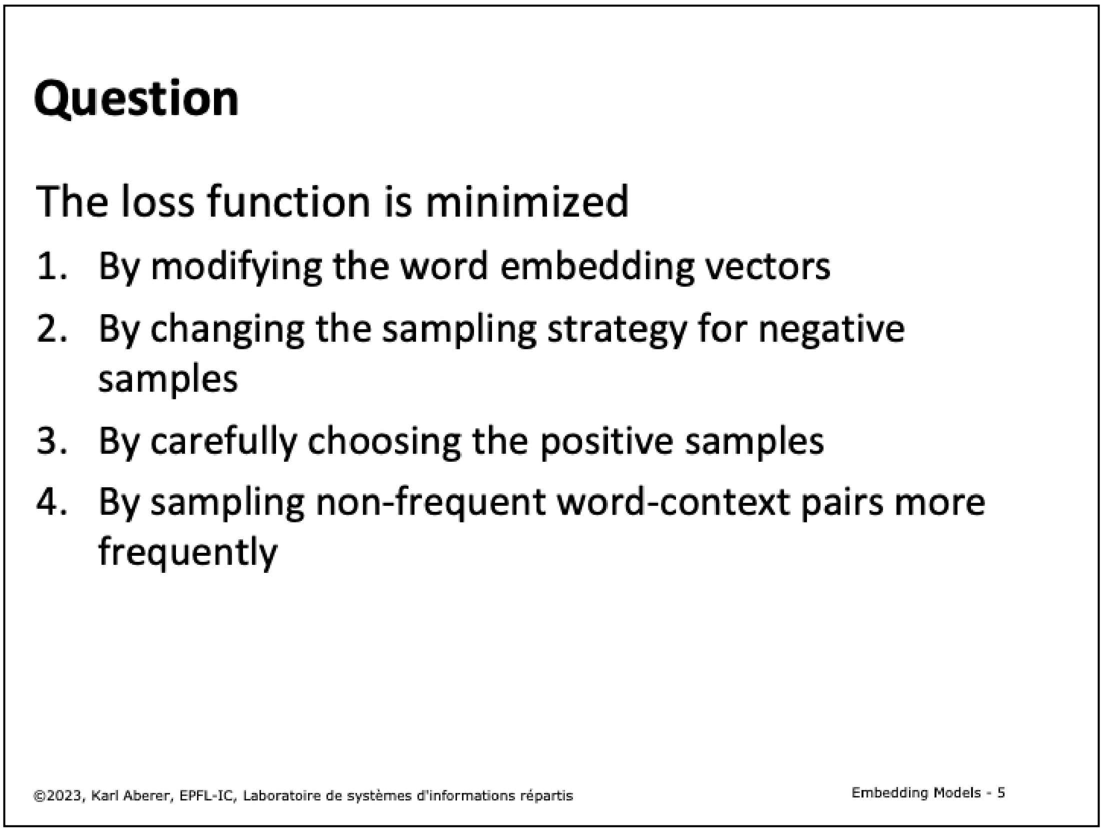
Question

With negative sampling a set of negative samples is created for

1. For each word of the vocabulary
2. For each word-context pair
3. For each occurrence of a word in the text
4. For each occurrence of a word-context pair in the text

Answer 4

For each occurrence of a word-context pair, a set of negative samples is produced. Note that this is also different from creating a set of negative samples for each word-context pair, since the same word-context pair can occur multiple times in the document collection.



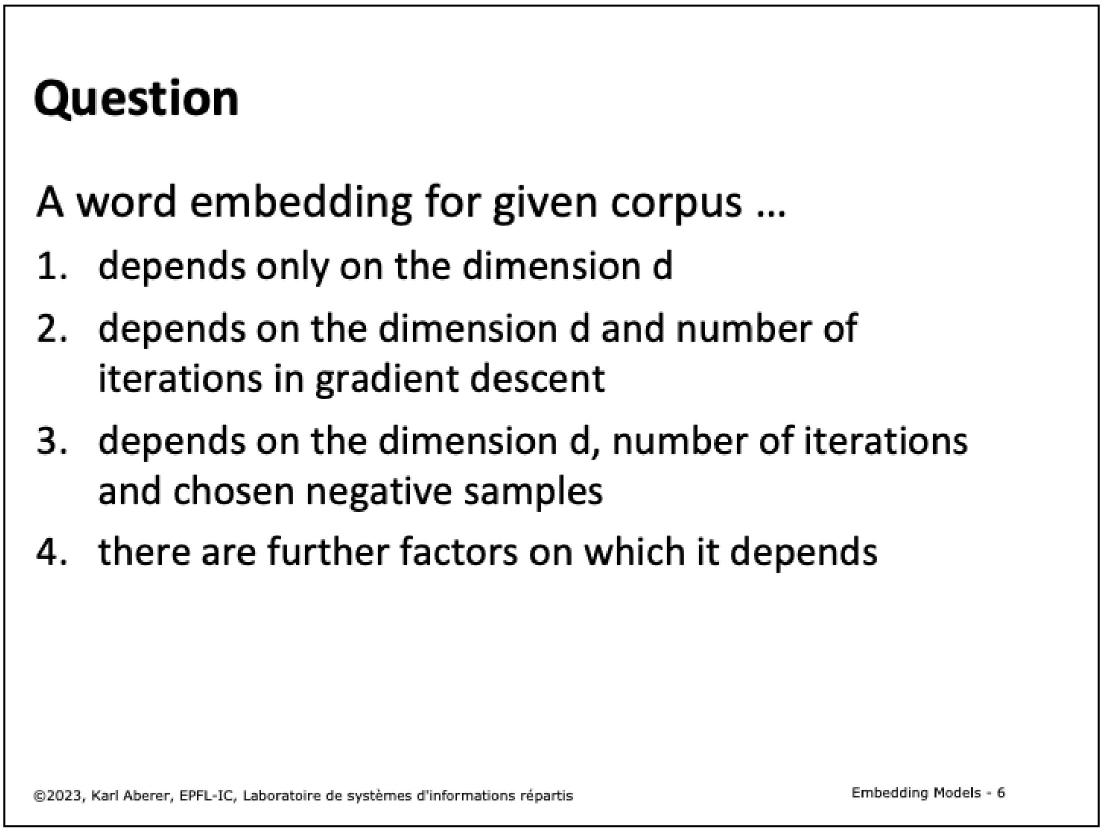
Question

The loss function is minimized

1. By modifying the word embedding vectors
2. By changing the sampling strategy for negative samples
3. By carefully choosing the positive samples
4. By sampling non-frequent word-context pairs more frequently

Answer 1

The loss function is minimized by incrementally modifying the word embedding vectors. Answer 4 refers to an approach to improve the quality of the word embeddings achieved, but not to minimize the loss function.



Question

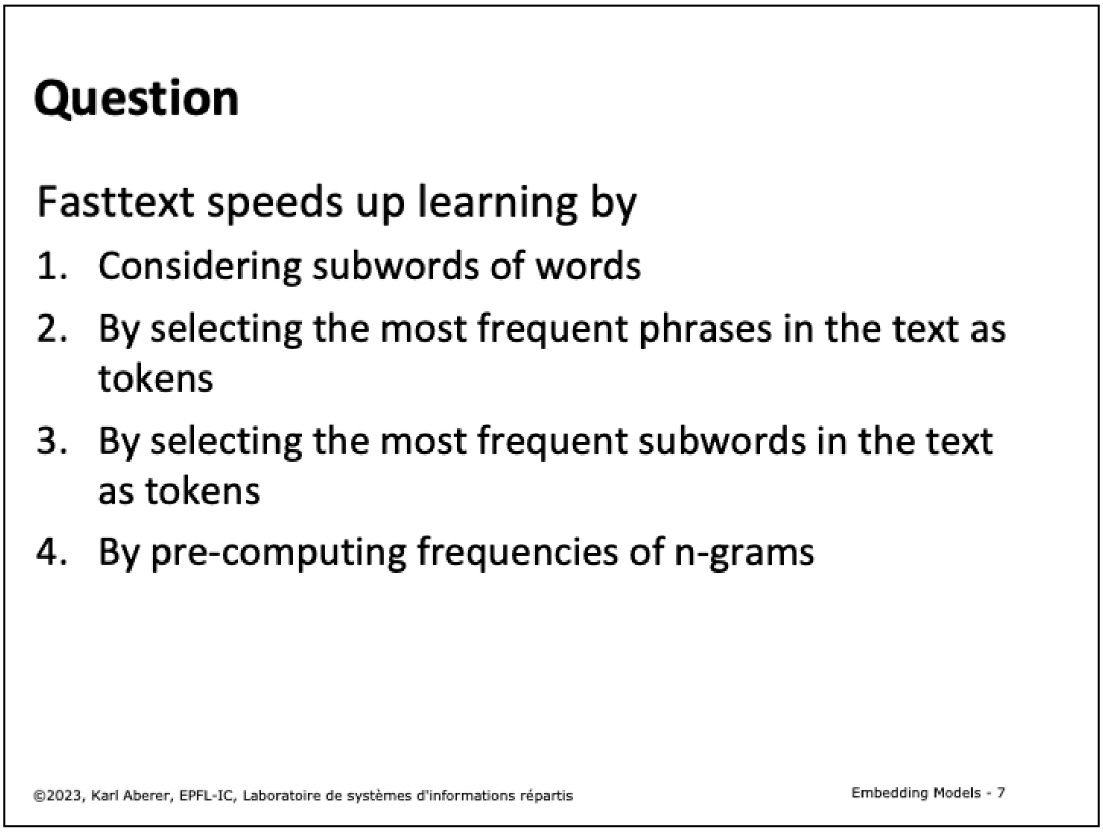
A word embedding for given corpus …

1. depends only on the dimension d
2. depends on the dimension d and number of iterations in gradient descent
3. depends on the dimension d, number of iterations and chosen negative samples
4. there are further factors on which it depends

Answer 4

Other factors that can influence the outcome of the optimization are

- The order in which the documents are processed
- The initialization of the word embedding vectors



Question

Fasttext speeds up learning by

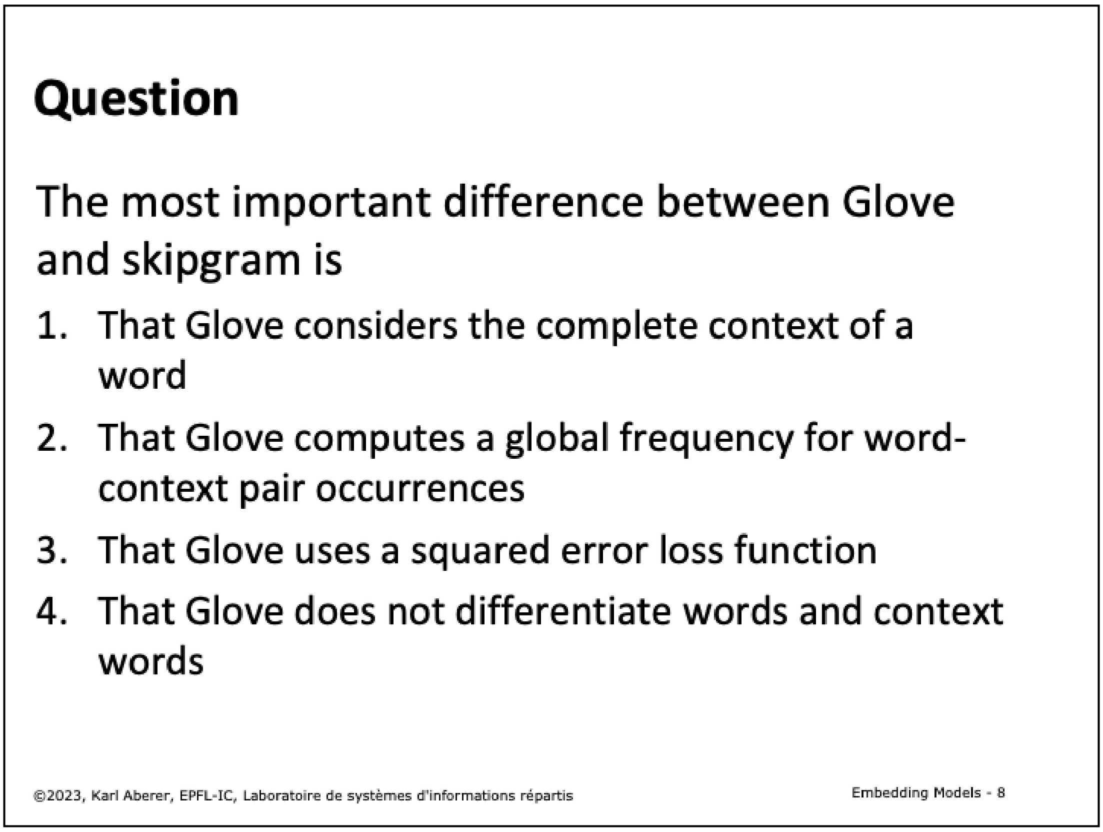
1. Considering subwords of words
2. By selecting the most frequent phrases in the text as tokens
3. By selecting the most frequent subwords in the text as tokens
4. By pre-computing frequencies of n-grams

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Embedding Models - 7

Answer 3

Answer 1 improves the quality of embeddings, but may slow down learning.
Answer 4 speeds up the selection of frequent phrases, but not learning.



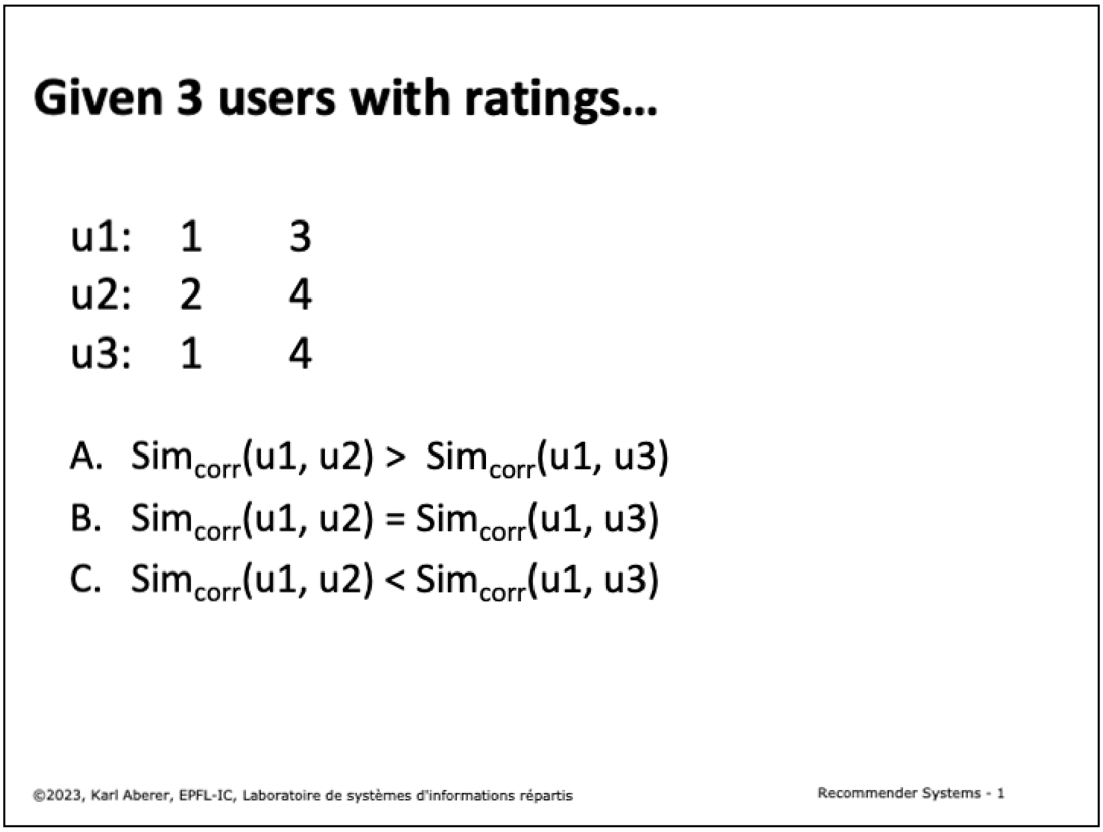
Question

The most important difference between Glove and skipgram is

1. That Glove considers the complete context of a word
2. That Glove computes a global frequency for word-context pair occurrences
3. That Glove uses a squared error loss function
4. That Glove does not differentiate words and context words

Answer 2

The main difference between Glove and earlier methods, including skipgram and CBOW, is the computation of global co-occurrence counts. This is an additional processing step, but provides additional information on the global statistics. Answer 3 refers to a difference that is rather a consequence of using a global statistics. As for answer 4, in a sense also skipgram does not really distinguish between words and context words.



Given 3 users with ratings…

u1:	1	3
u2:	2	4
u3:	1	4

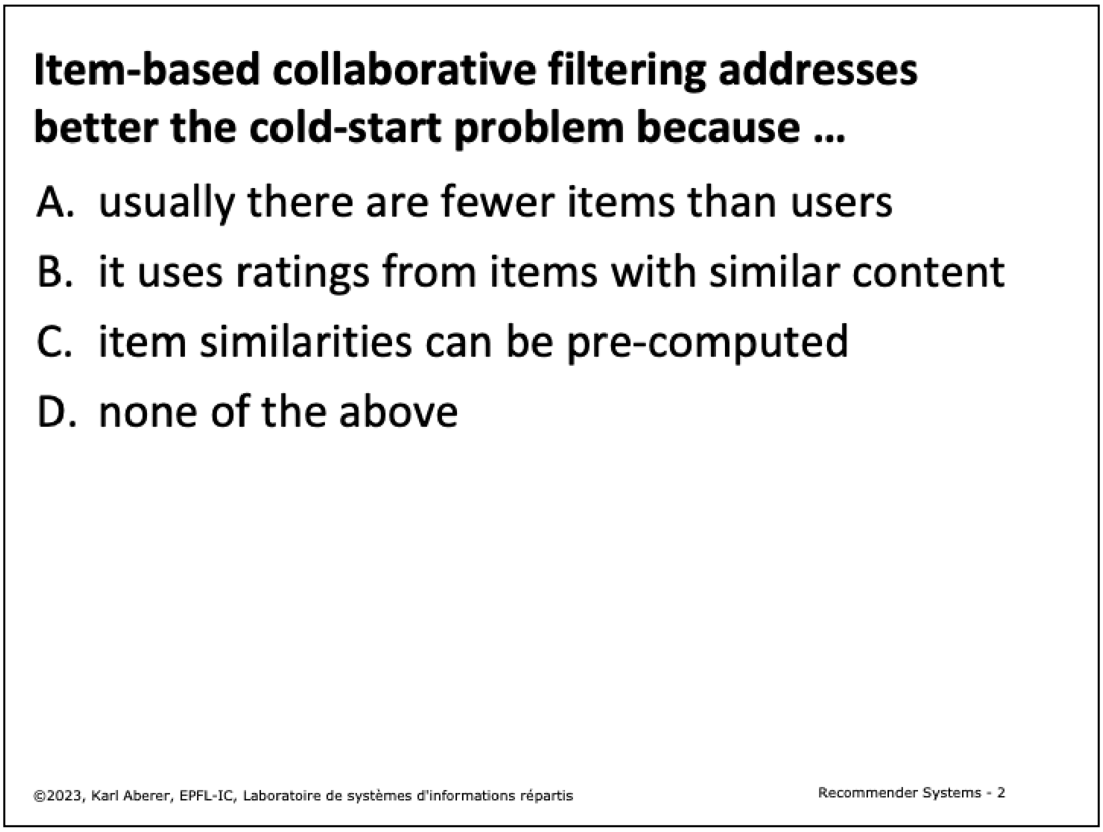
A. $\text{Sim}_{corr}(u1, u2) > \text{Sim}_{corr}(u1, u3)$

B. $\text{Sim}_{corr}(u1, u2) = \text{Sim}_{corr}(u1, u3)$

C. $\text{Sim}_{corr}(u1, u2) < \text{Sim}_{corr}(u1, u3)$

Answer B

With Pearson correlation, similarity is computed using vectors that are centered around the mean and the normalized. Therefore, after normalization all three vectors are the same, namely 1/sqrt(2) * (-1,1).



Item-based collaborative filtering addresses better the cold-start problem because …

A. usually there are fewer items than users

B. it uses ratings from items with similar content

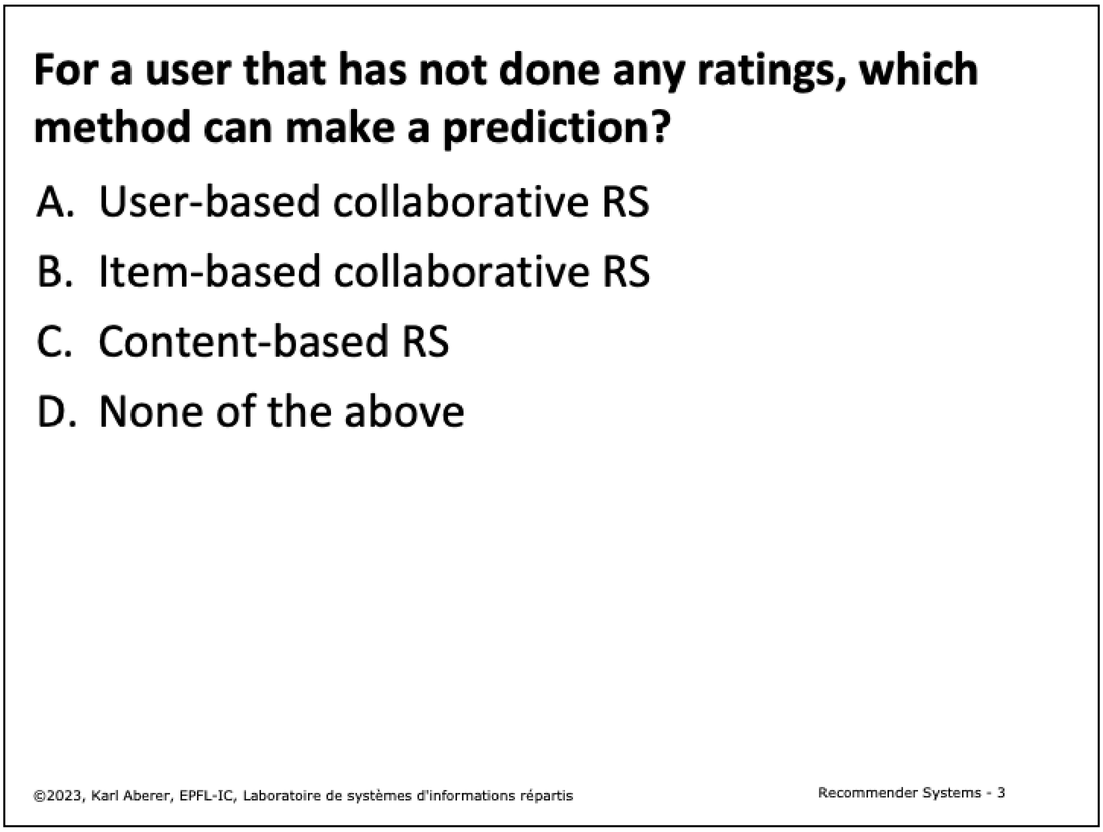
C. item similarities can be pre-computed

D. none of the above

Answer D

If a user has no ratings, also item-based collaborative filtering cannot make a prediction on the user’s rating of an unknown item.

If the item in question has not ratings, not similar items can be determined and no predication can be made as well.



For a user that has not done any ratings, which method can make a prediction?

A. User-based collaborative RS

B. Item-based collaborative RS

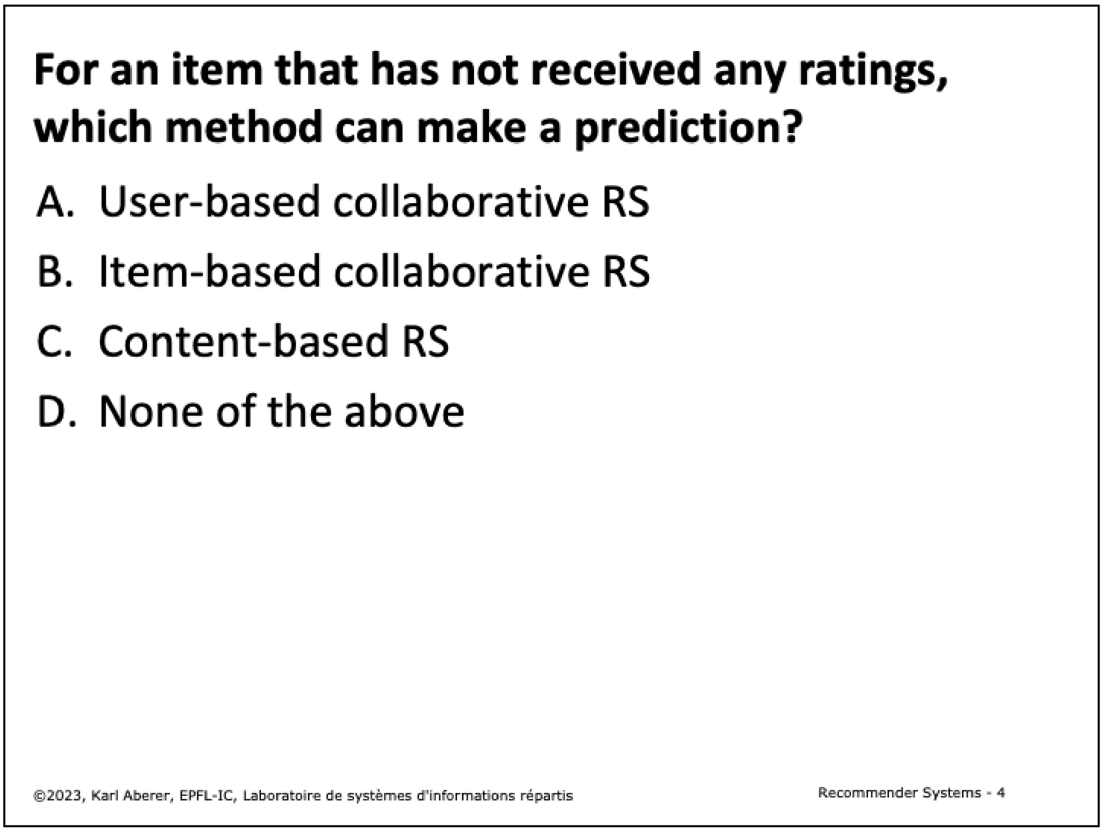
C. Content-based RS

D. None of the above

Answer D

User-base and item-based collaborative filtering cannot make predictions without previous ratings of the users.

Since the user has not rated any item, also content-based filtering is not applicable.



For an item that has not received any ratings, which method can make a prediction?

A. User-based collaborative RS

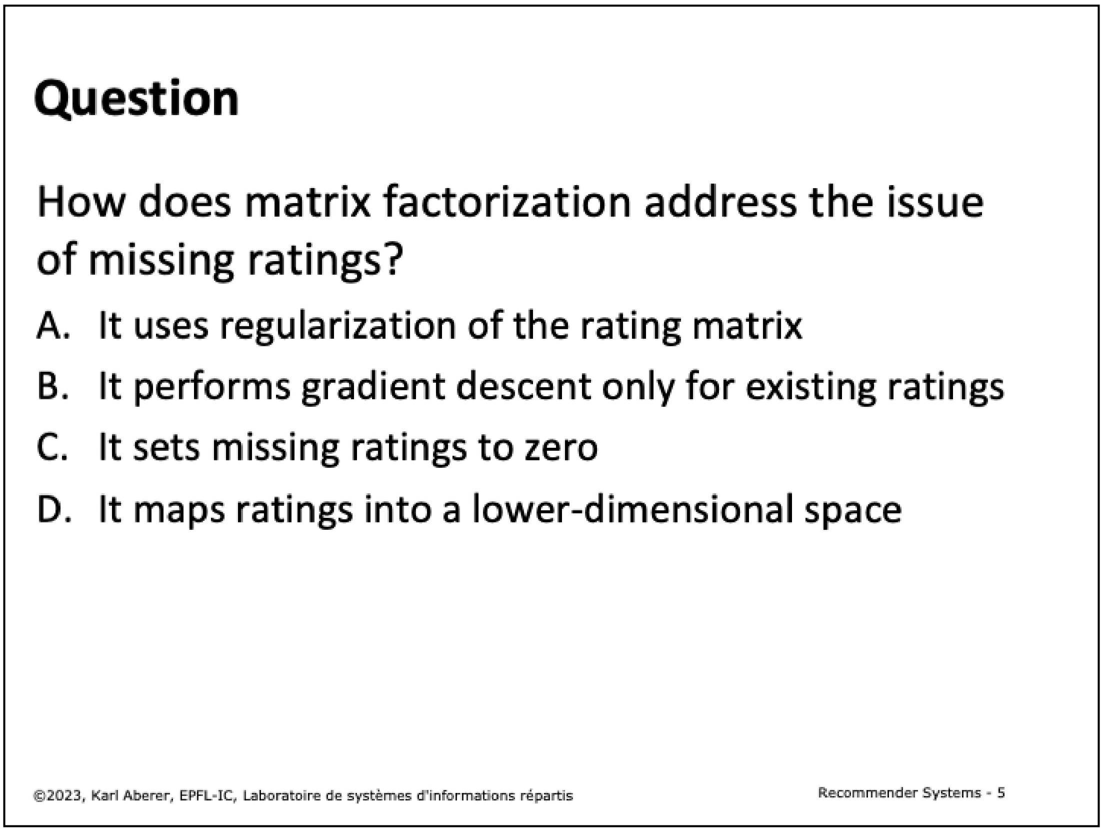
B. Item-based collaborative RS

C. Content-based RS

D. None of the above

Answer C

Based on the content, similat items can be found, even if the item has not yet received a rating.



Question

How does matrix factorization address the issue of missing ratings?

A. It uses regularization of the rating matrix

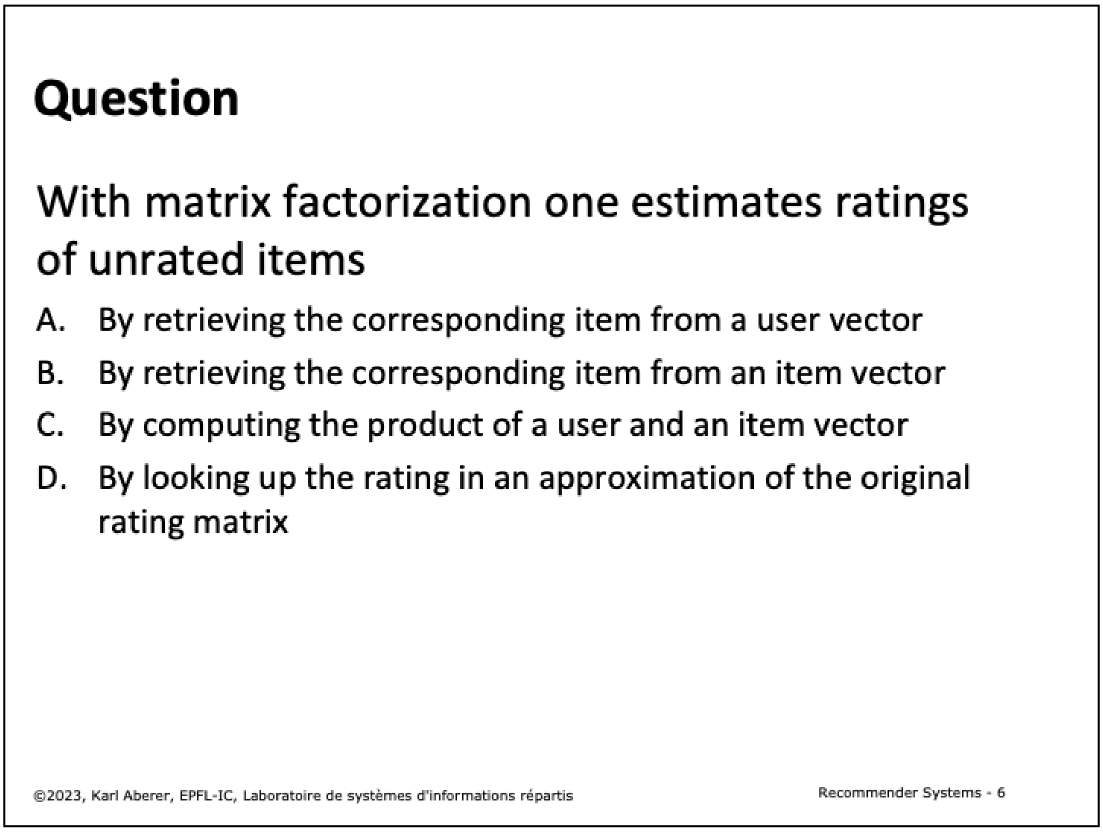
B. It performs gradient descent only for existing ratings

C. It sets missing ratings to zero

D. It maps ratings into a lower-dimensional space

Answer B

In gradient descent only data on existing ratings (user-item pairs) are exploited.



Question

With matrix factorization one estimates ratings of unrated items

A. By retrieving the corresponding item from a user vector

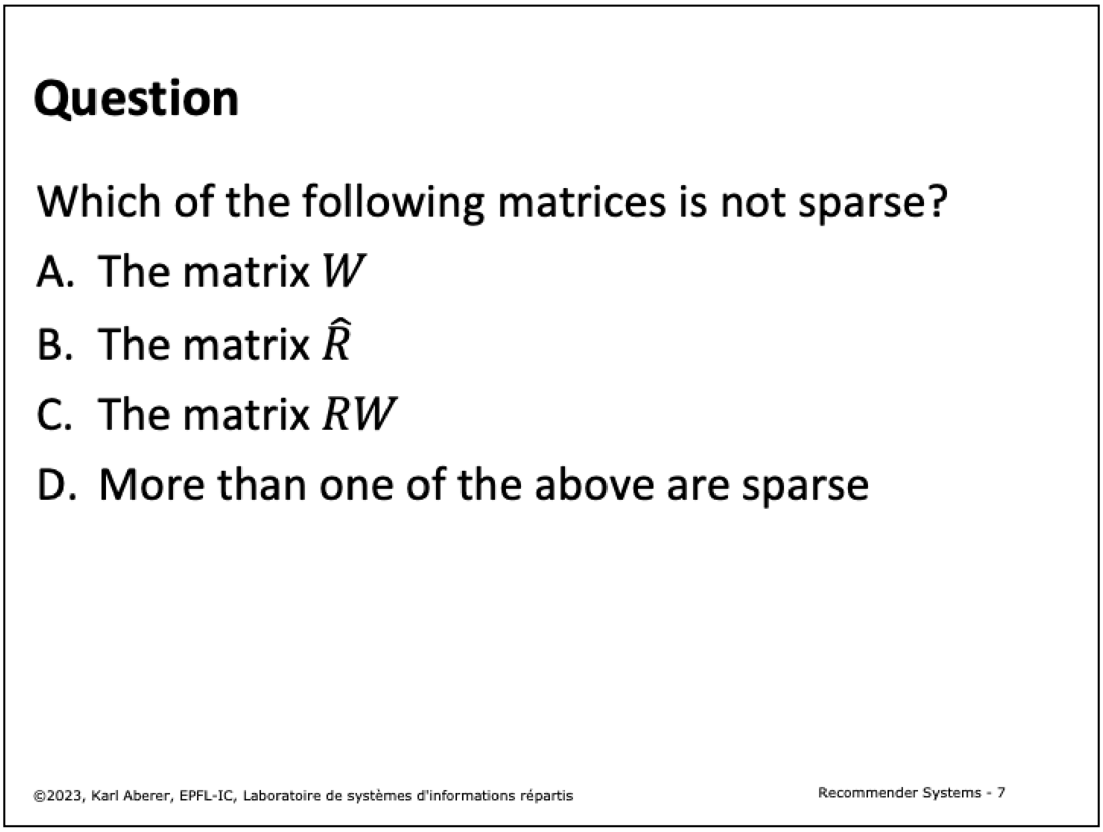
B. By retrieving the corresponding item from an item vector

C. By computing the product of a user and an item vector

D. By looking up the rating in an approximation of the original rating matrix

Answer C

For estimating a new rating the latent representations of the corresponding user and item are used to compute an estimation. In principle, one could also consider Answer D as correct, but this would imply that all ratings are precomputed, which would be a waste of resources, if only few ratings need to be known.



Question

Which of the following matrices is not sparse?

A. The matrix W

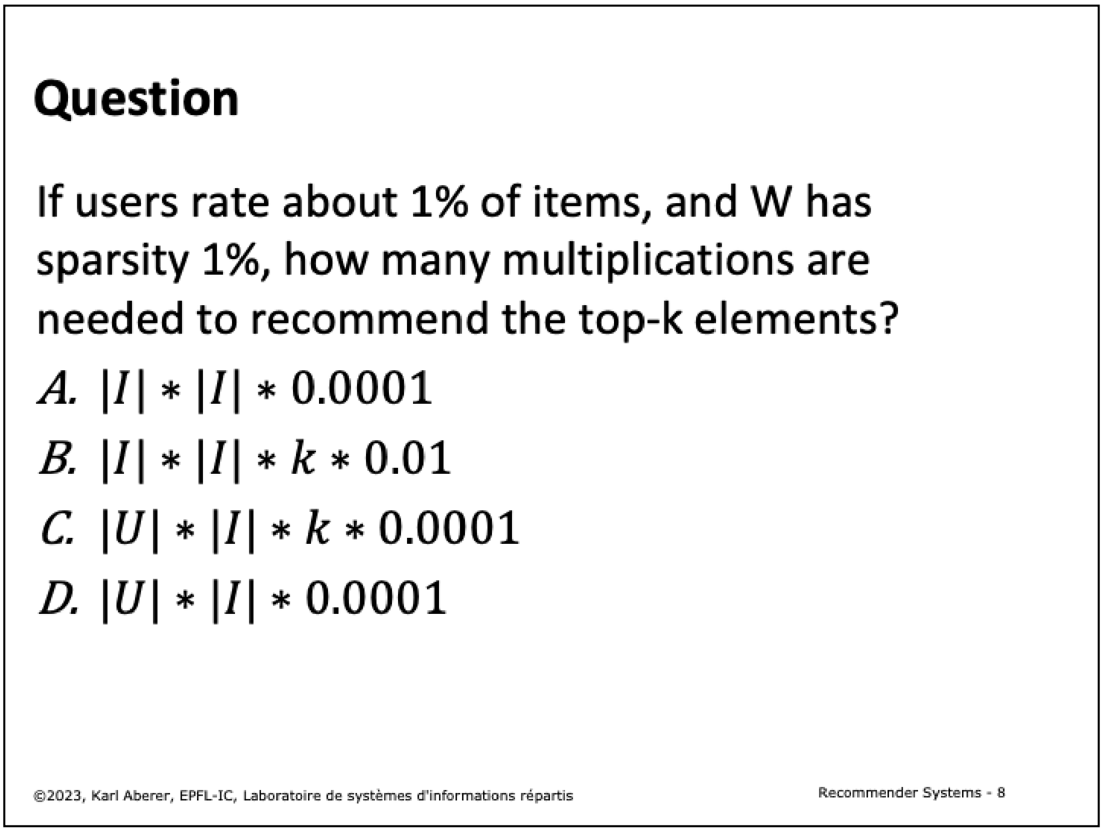
B. The matrix $\hat{R}$

C. The matrix RW

D. More than one of the above are sparse

Answer A

Only the matrix W is sparse. The two other matrices are the same, and contain rating estimations for all items. Therefore, they are not sparse.



Question

If users rate about 1% of items, and W has sparsity 1%, how many multiplications are needed to recommend the top-k elements?

A. $|I| * |I| * 0.0001$

B. $|I| * |I| * k * 0.01$

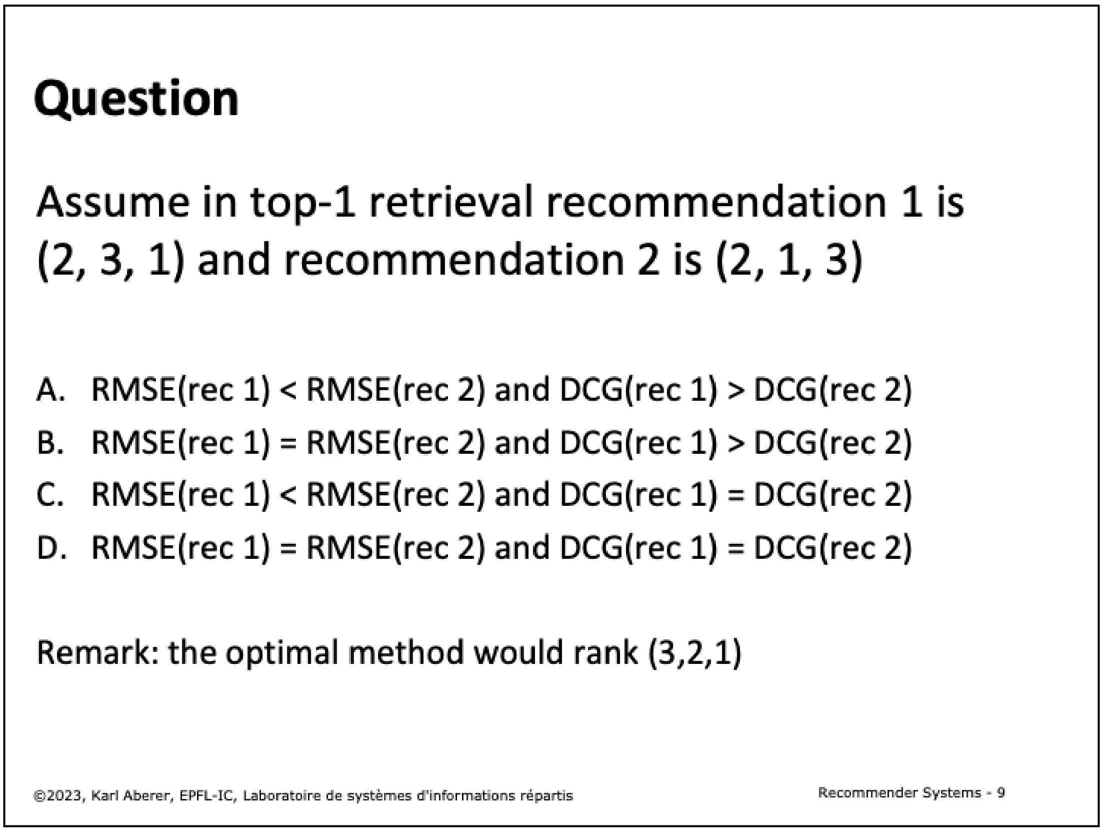
C. $|U| * |I| * k * 0.0001$

D. $|U| * |I| * 0.0001$

Recommender Systems - 8

Answer A

For finding the top k ratings, for each non-rated item of the user (about |I|), the rating vector of the user has to be multiplied with the corresponding item vector in W. The user rating vector has about |I| * 0.01 non-zero entries, as well as the item vector. Computing the scalar product of a dense vector takes |I| multiplictaions. Therefore, in total |I|^2 * 0.00001 multiplications will be performed.



Question

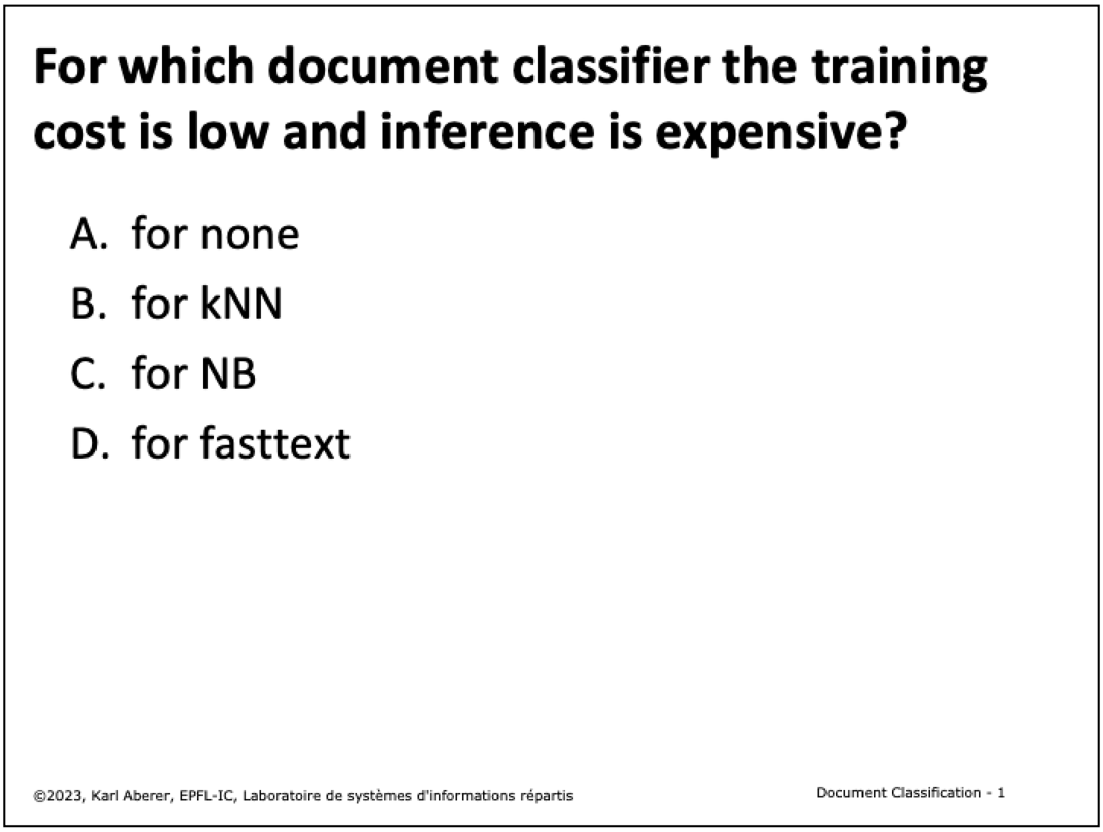
Assume in top-1 retrieval recommendation 1 is (2, 3, 1) and recommendation 2 is (2, 1, 3)

A. RMSE(rec 1) < RMSE(rec 2) and DCG(rec 1) > DCG(rec 2)
B. RMSE(rec 1) = RMSE(rec 2) and DCG(rec 1) > DCG(rec 2)
C. RMSE(rec 1) < RMSE(rec 2) and DCG(rec 1) = DCG(rec 2)
D. RMSE(rec 1) = RMSE(rec 2) and DCG(rec 1) = DCG(rec 2)

Remark: the optimal method would rank (3,2,1)

Answer C

Since DCG considers only the top-k items, in this case the first item, the measure is independent of the ordering of the non top-k items. RMSE will give a better score to rec 1, since the higher rated item 3 is ranked higher in the result.



For which document classifier the training cost is low and inference is expensive?

A. for none

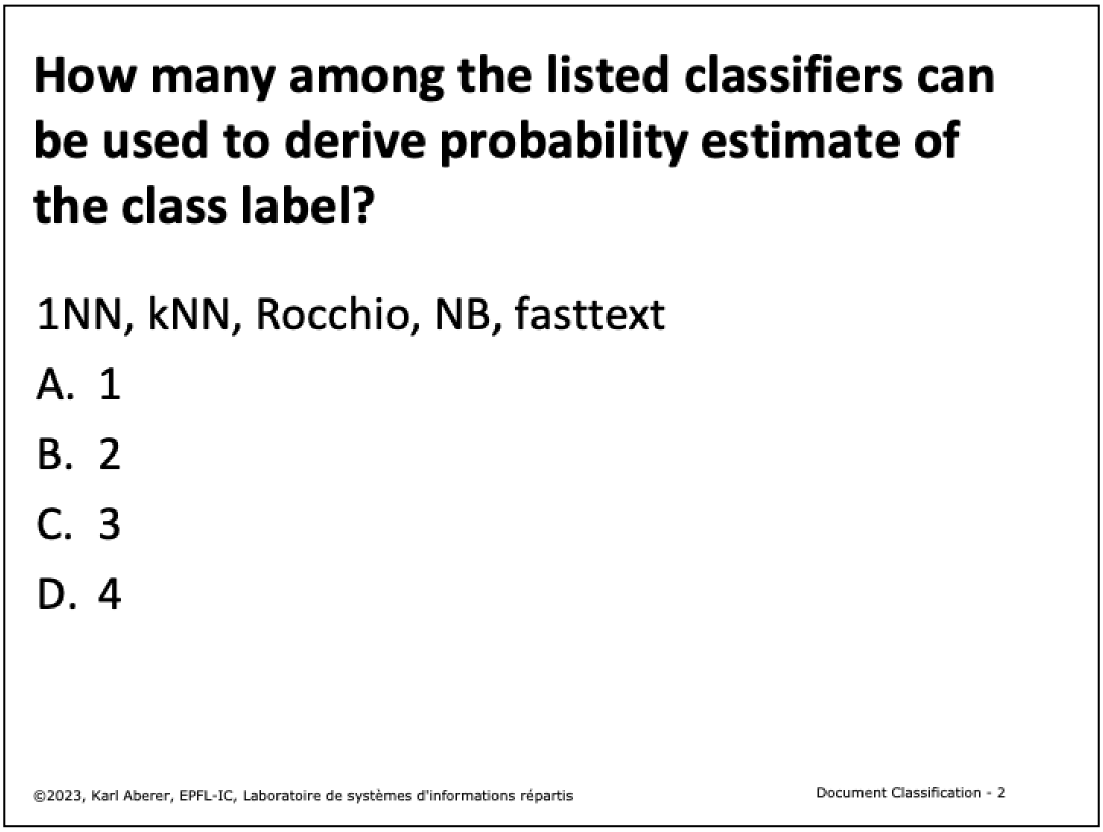
B. for kNN

C. for NB

D. for fasttext

Answer B

For kNN the only cost incurred for training is the computation of document weights, including the idf values. In Naïve Bayes for all words the likelihood of the word to occur in a class has to be computed. Fasttext requires SGD.



How many among the listed classifiers can be used to derive probability estimate of the class label?

1NN, kNN, Rocchio, NB, fasttext

A. 1

B. 2

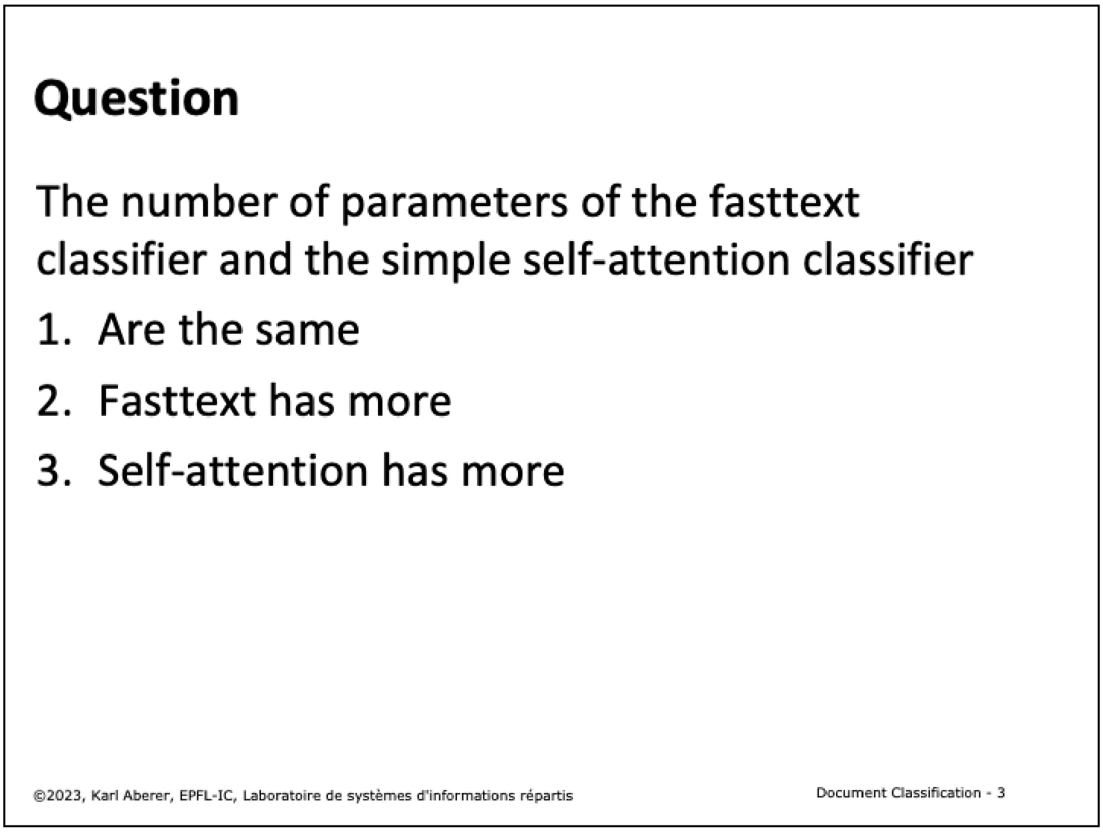
C. 3

D. 4

Answer C

1NN cannot be used to compute a probability estimate, since it provides a binary decision of which is the closest class label. The same is true for the Rocchio method.

kNN an be used to estimate the probabilities for class labels if k is sufficiiently large. NB and fasttext by defintion provide probability estimates.



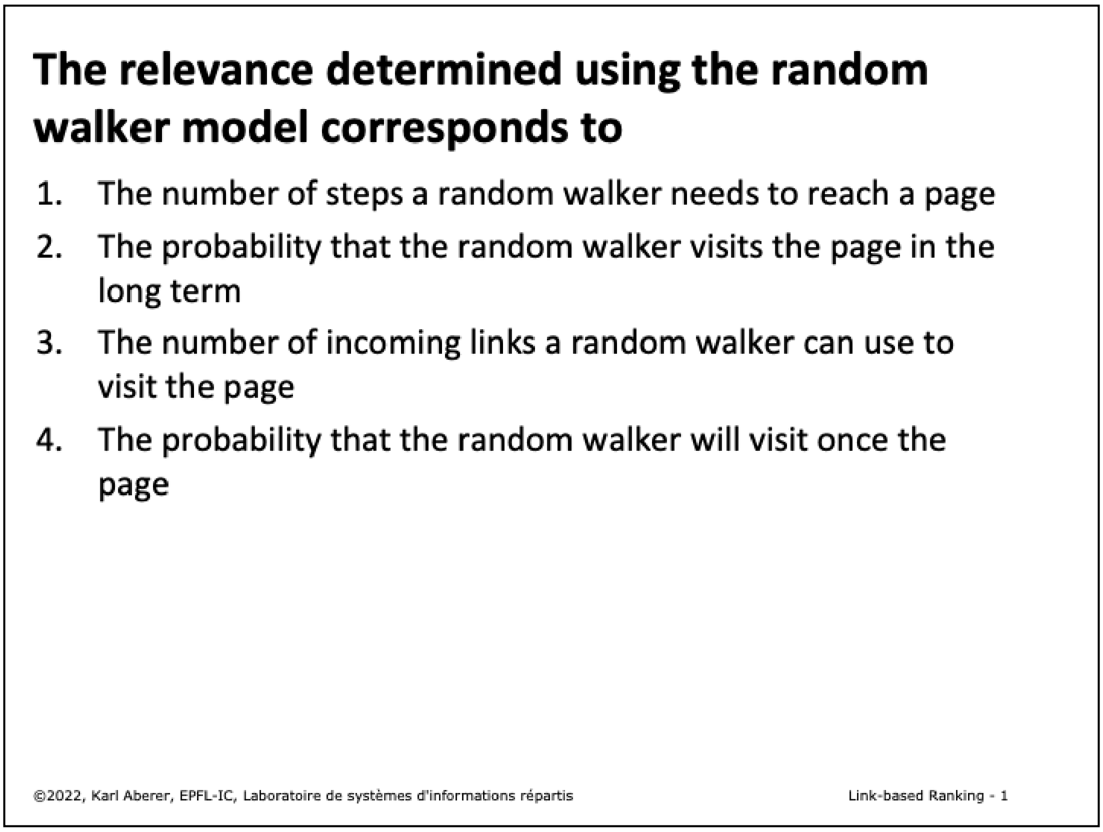
Question

The number of parameters of the fasttext classifier and the simple self-attention classifier

1. Are the same
2. Fasttext has more
3. Self-attention has more

Answer 1

The simple self-attention based classifier applied before the averaging of the word vectors the self-attention mechanism. This mechanism uses only the word vectors as weights and therefore does nto introduce new parameters into the model.



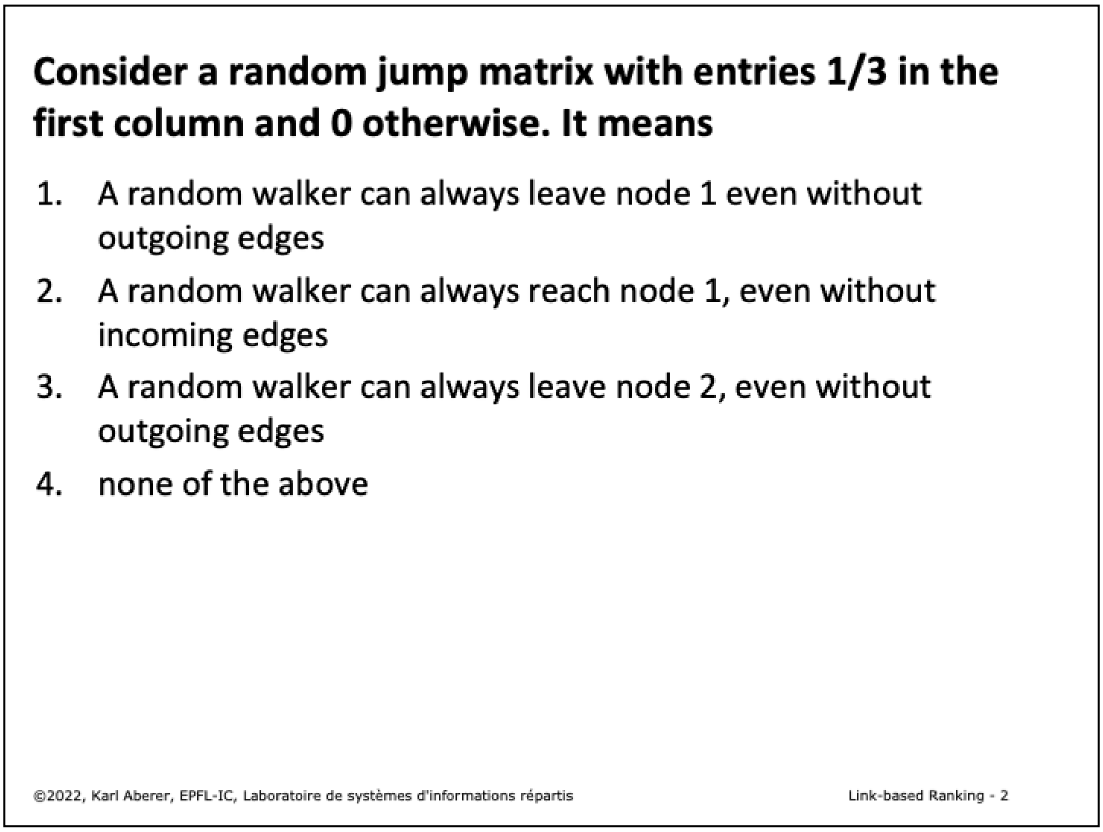
The relevance determined using the random walker model corresponds to

1. The number of steps a random walker needs to reach a page
2. The probability that the random walker visits the page in the long term
3. The number of incoming links a random walker can use to visit the page
4. The probability that the random walker will visit once the page

Link-based Ranking - 1

Answer 2

The relevance measure derived from the random walker model corresponds to the stationary (long-term) probability of the random walker to visit a page. The number of incoming links does have an influence on this probability but is not determining it.



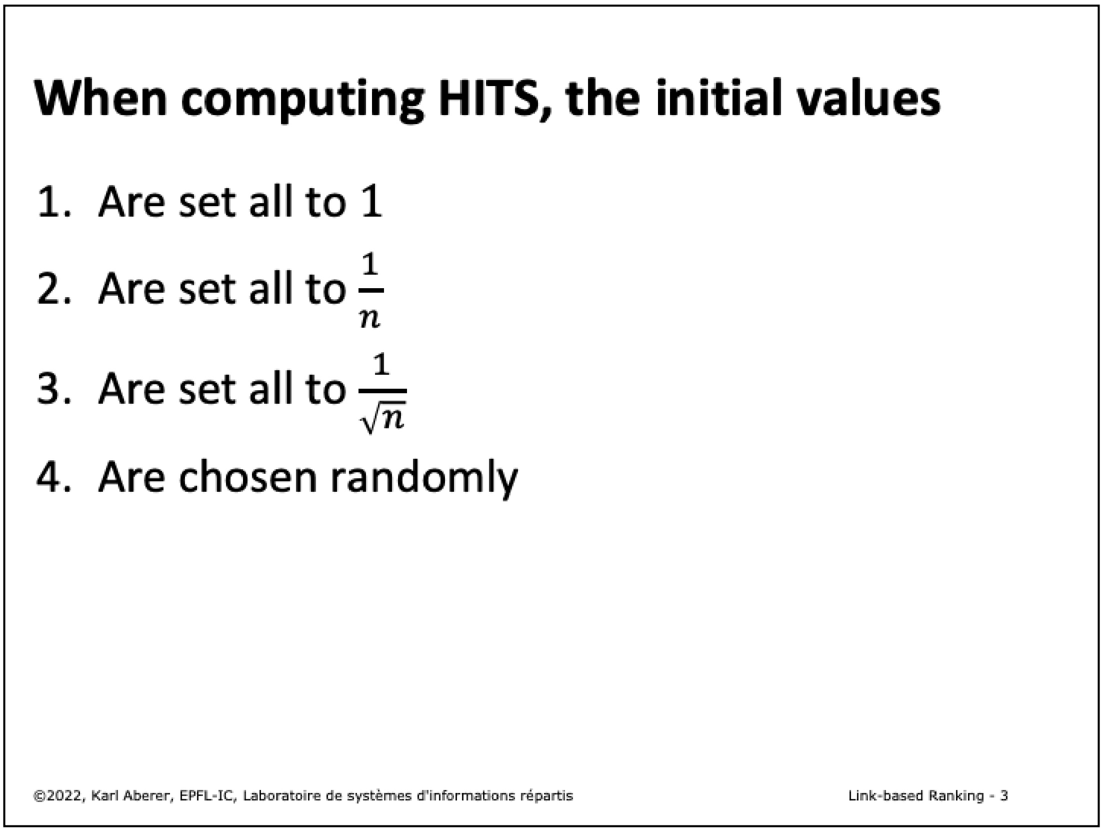
Consider a random jump matrix with entries 1/3 in the first column and 0 otherwise. It means

1. A random walker can always leave node 1 even without outgoing edges
2. A random walker can always reach node 1, even without incoming edges
3. A random walker can always leave node 2, even without outgoing edges
4. none of the above

©2022, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis Link-based Ranking - 2

Answer 1

The first column of the matrix corresponds to the probabilities to jump to from node 1 to every other node. Therefore, the random walker when arriving at node 1 can always leave that node by making random jump. On the other hand, node 1 can only be reached when an incoming edge exists. If node 2 has no outgoing edge, the random walker cannot leave it.

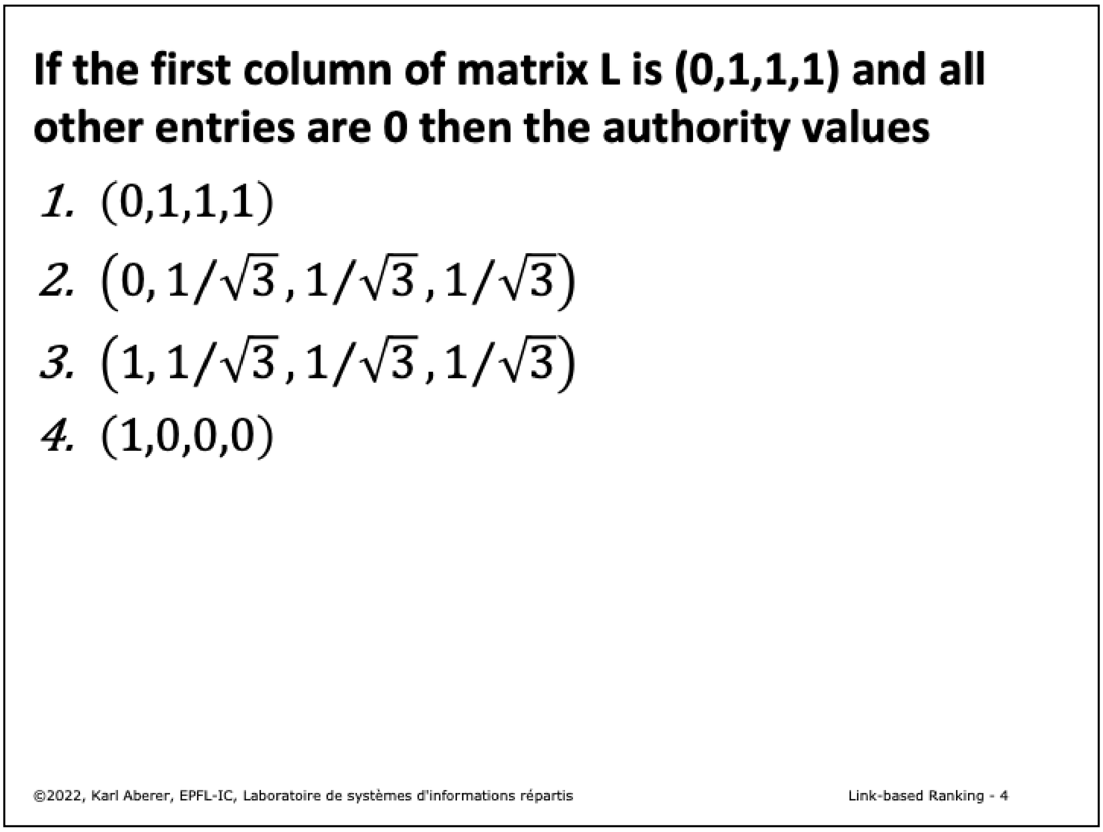


When computing HITS, the initial values

1. Are set all to 1
2. Are set all to $\frac{1}{n}$
3. Are set all to $\frac{1}{\sqrt{n}}$
4. Are chosen randomly

Answer 3

In the standard formulation of the algorithm the normalization factor is $\frac{1}{\sqrt{n}}$ so that the L2 norm of the vector equals 1.



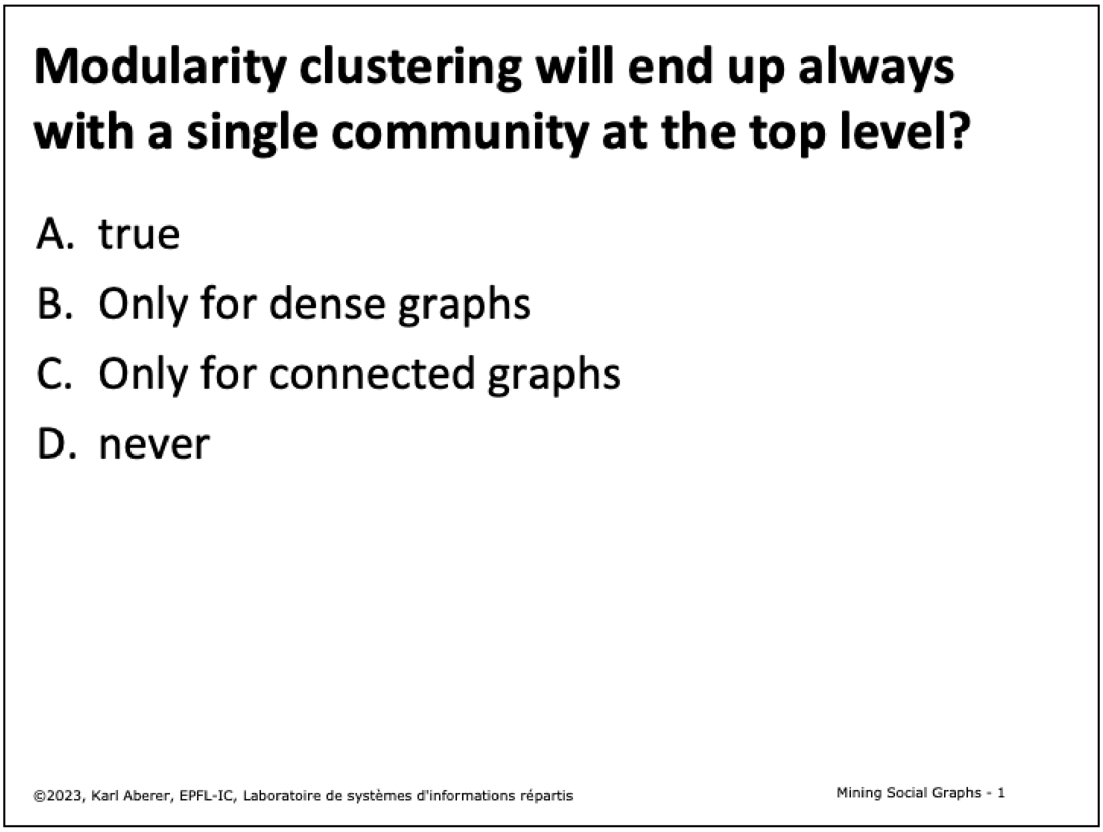
If the first column of matrix L is (0,1,1,1) and all other entries are 0 then the authority values

1. $(0,1,1,1)$
2. $\left(0, 1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}\right)$
3. $\left(1, 1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}\right)$
4. $(1,0,0,0)$

©2022, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis Link-based Ranking - 4

Answer 2

If the first column of matrix L is (0,1,1,1), then node 1 has one link to each other graph. Correspondingly it will be a hub, and the only hub and have hub weight 1. The other nodes will receive the same authority from that node. Thus, their weights will be equally distributed and be therefore $1/\sqrt{3}$ so that the L2 norm of the authority vector equals 1. Note that as consequence node 1 will receive equal weights from the three authority nodes.



Modularity clustering will end up always with a single community at the top level?

A. true

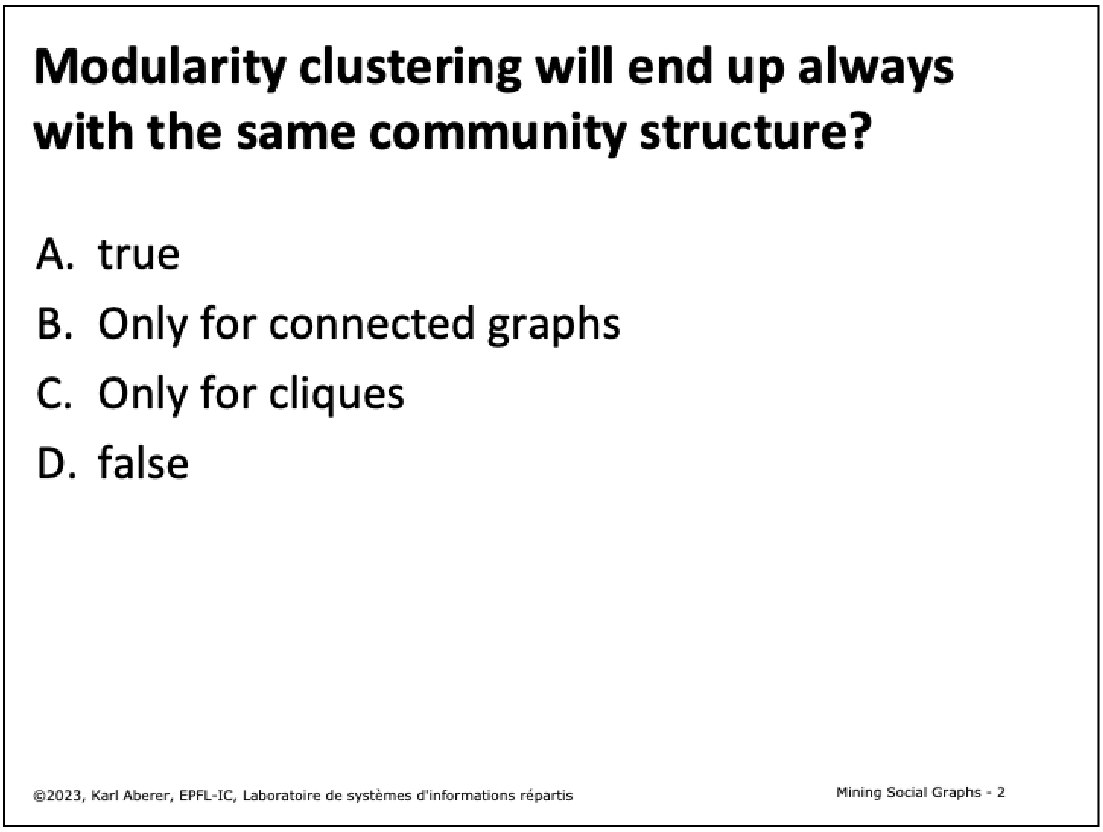
B. Only for dense graphs

C. Only for connected graphs

D. never

Answer C

If the graph is disconnected in distinct subgraphs, modularity will end up with a node representing each of those subgraphs that are not connected to each other and will therefore not be grouped to a community.



Modularity clustering will end up always with the same community structure?

A. true

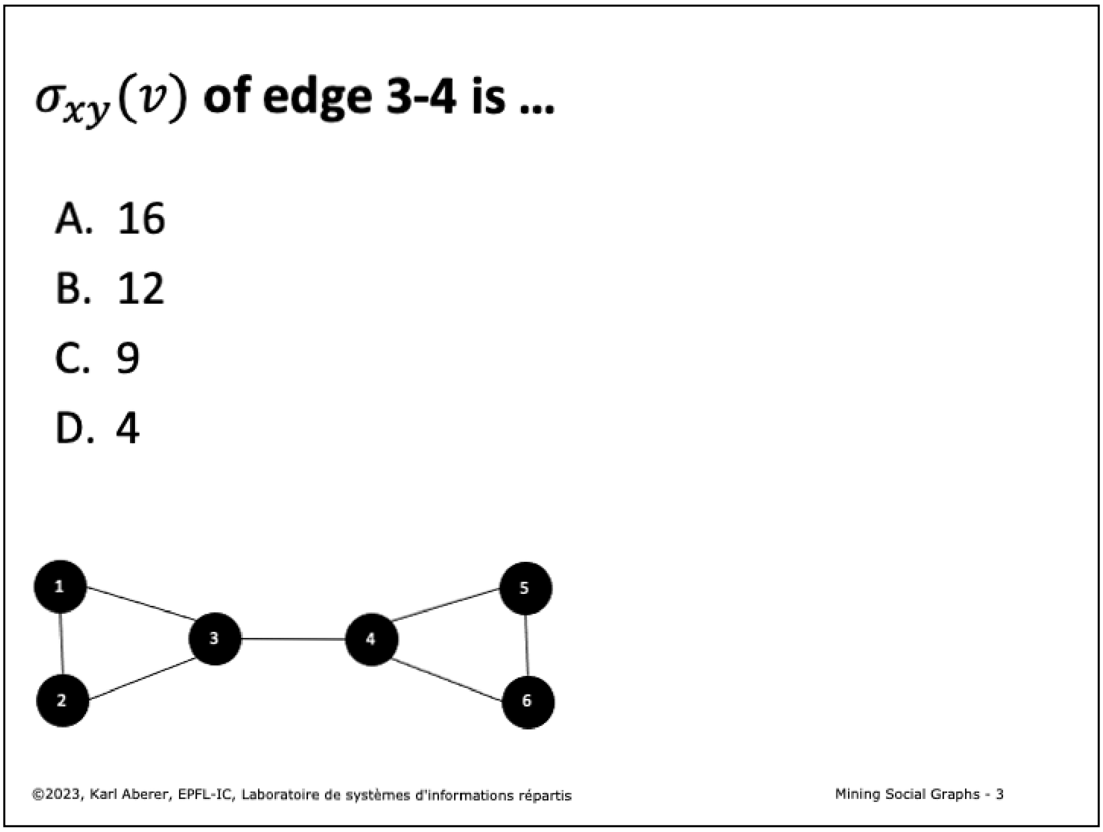
B. Only for connected graphs

C. Only for cliques

D. false

Answer D

They detailed outcome of modularity clustering depends on the order of processing of the nodes. By changing the processing order different communities may result. For cliques always one community will be formed in one pass (to be shown), but other regular graph structures, e.g. a ring, may also have this property.



$\sigma_{xy}(v)$ **of edge 3-4 is …**

A. 16

B. 12

C. 9

D. 4

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

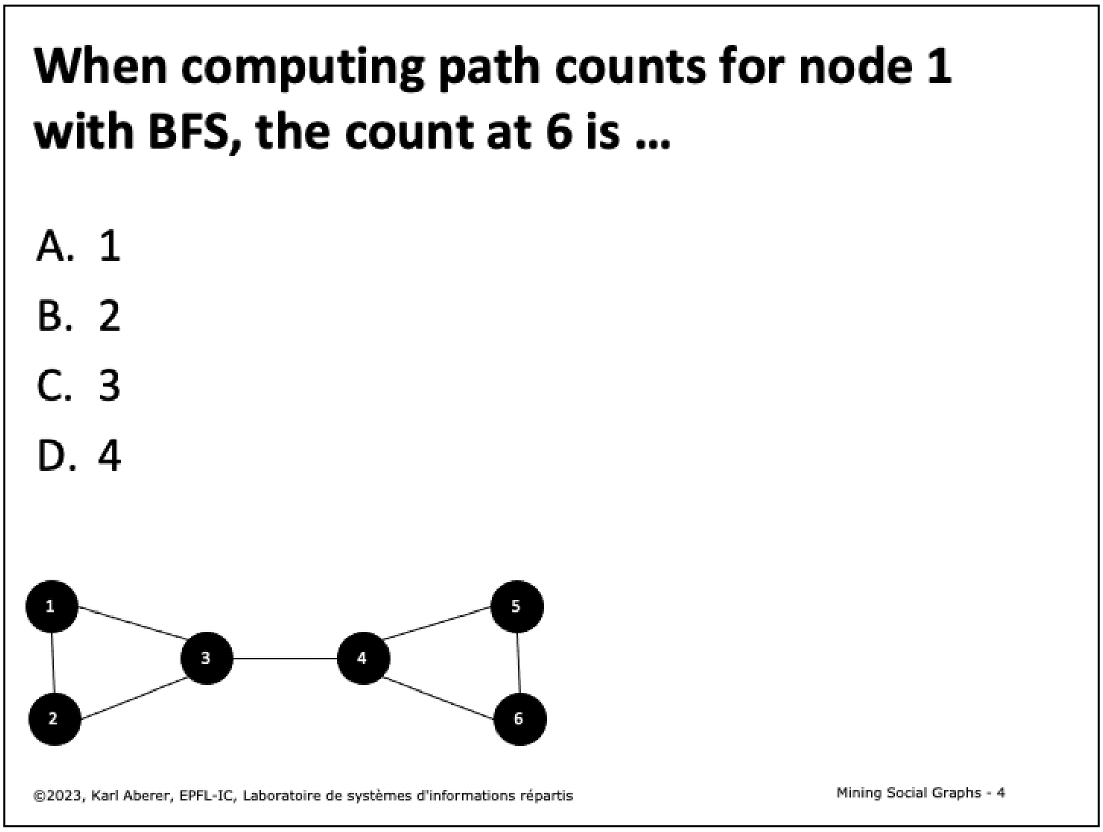
Mining Social Graphs - 3

Answer C

We have 4 shortest paths from {1,2} to {5, 6} passing through edge 3-4

In addition, there are 2 shortest paths from 3 to {5, 6} and from 4 to {1, 2} and the path from 3 to 4.

This gives total 9 paths.



When computing path counts for node 1 with BFS, the count at 6 is …

A. 1

B. 2

C. 3

D. 4

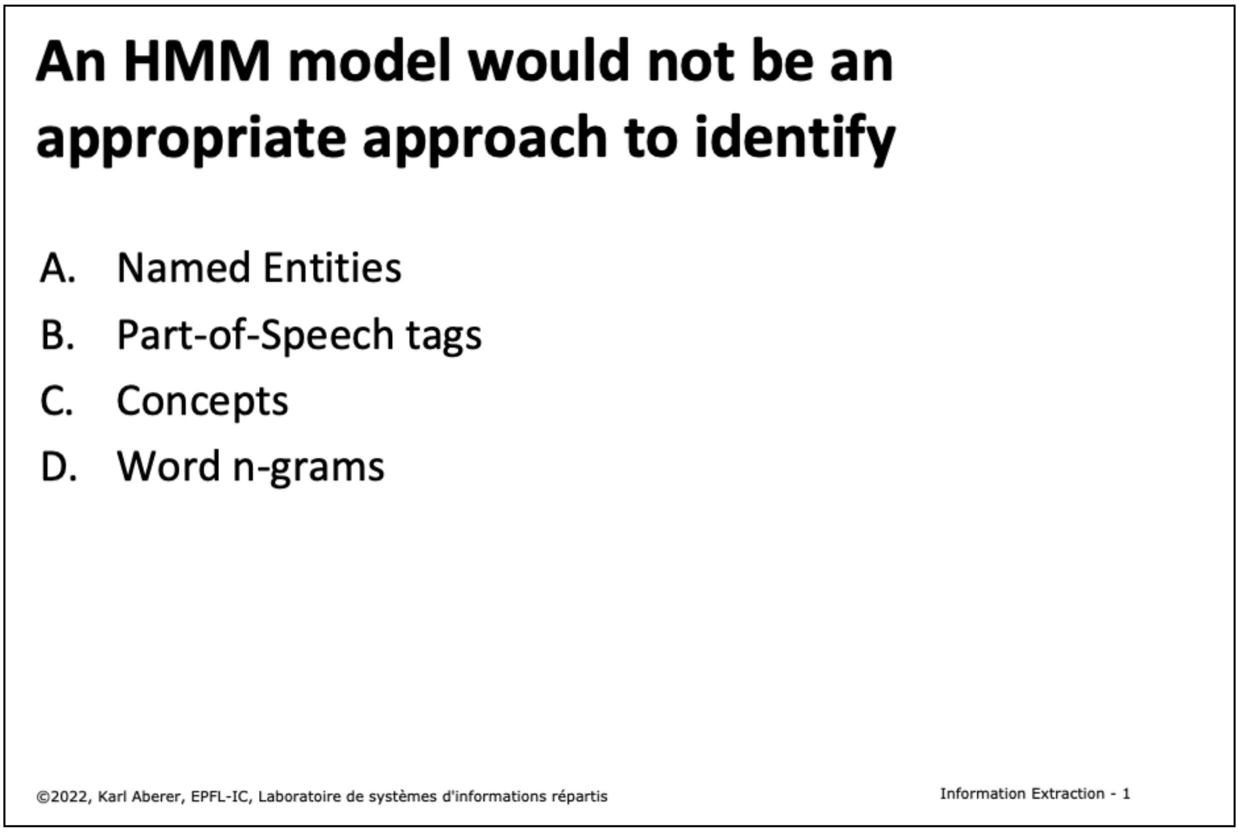
©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Mining Social Graphs - 4

Answer A

BFS computes the number of shortest paths from the starting node 1 to every other node of the graph.

Since there exists only one shortest path from 1 to 6, this count is 1.



An HMM model would not be an appropriate approach to identify

A. Named Entities

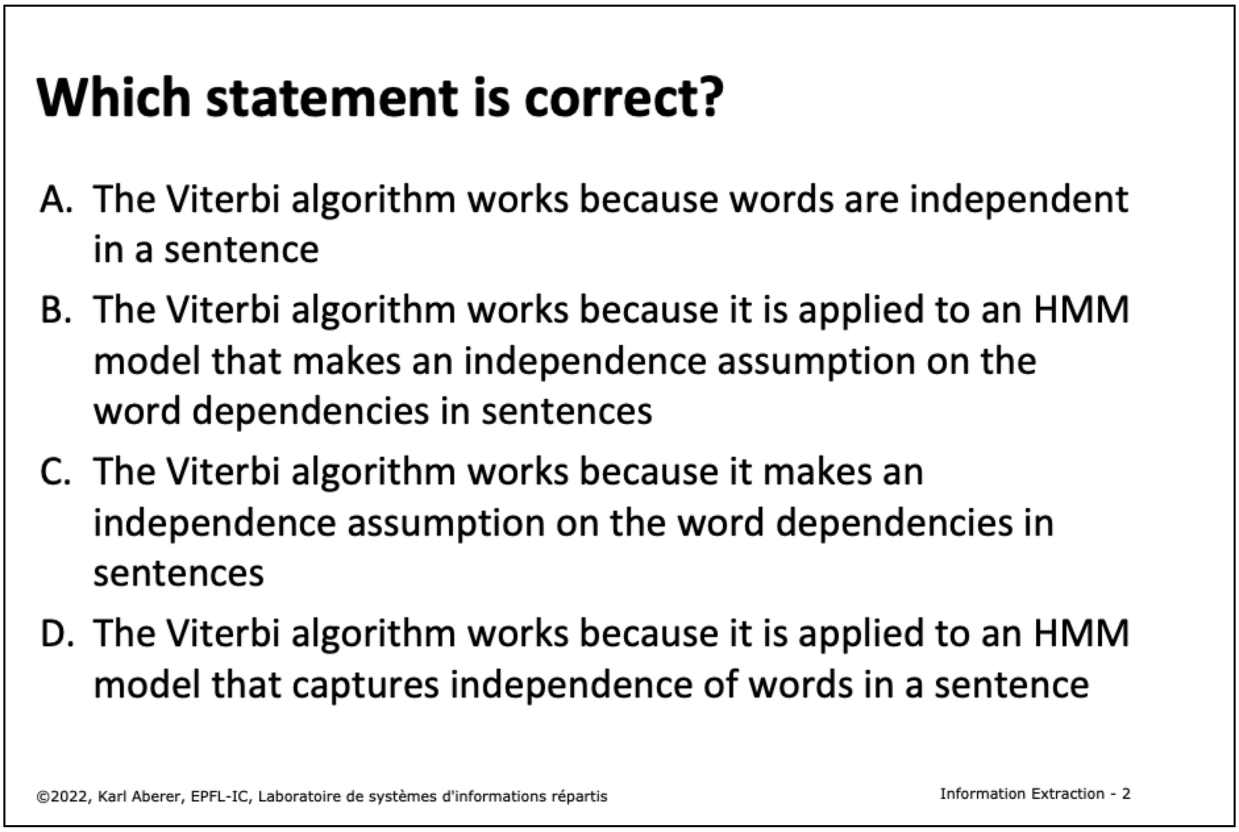
B. Part-of-Speech tags

C. Concepts

D. Word n-grams

Answer D

Word n-grams can be determined by simple algortihmic search. All other types of information are suitable to be modelled using HMMs.



Which statement is correct?

A. The Viterbi algorithm works because words are independent in a sentence

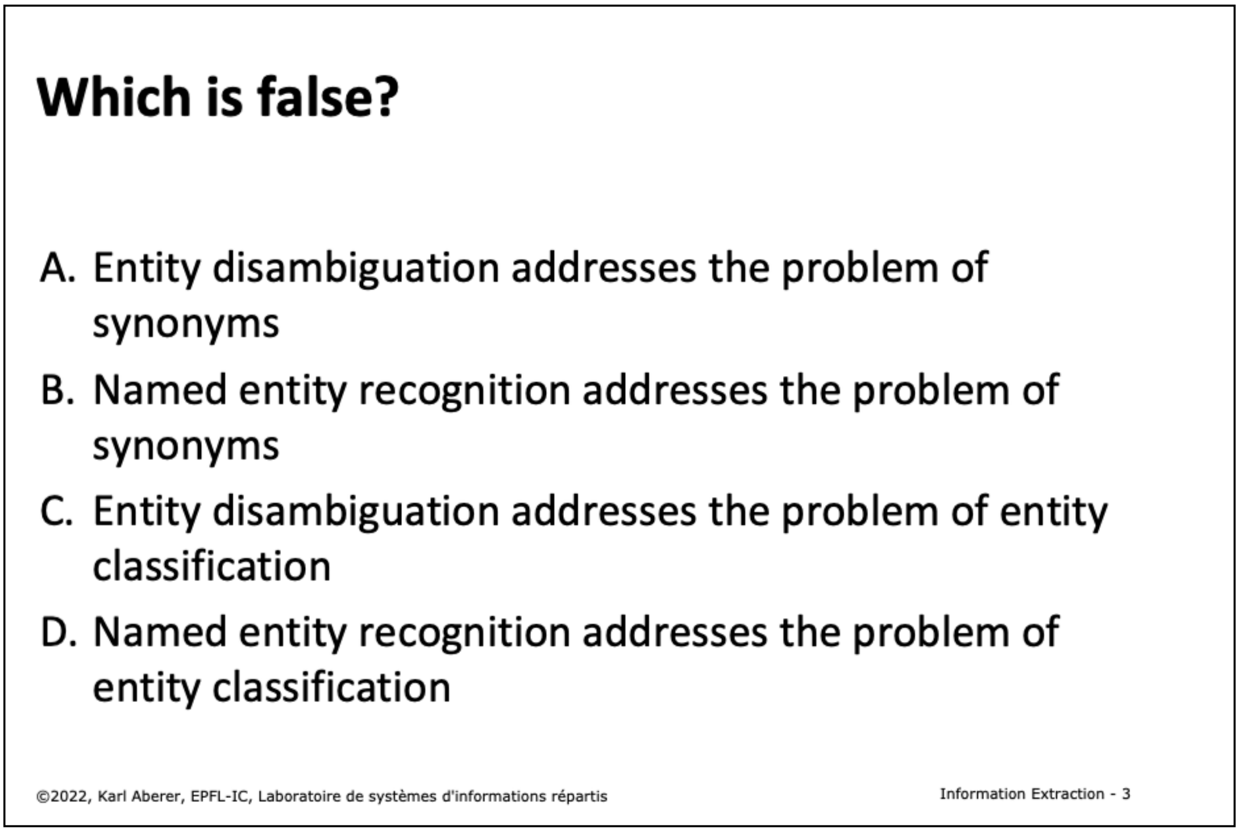
B. The Viterbi algorithm works because it is applied to an HMM model that makes an independence assumption on the word dependencies in sentences

C. The Viterbi algorithm works because it makes an independence assumption on the word dependencies in sentences

D. The Viterbi algorithm works because it is applied to an HMM model that captures independence of words in a sentence

Answer B

All other answers have some logical problem in the formulation. Answer A would imply that works are independent in a sentence, which is clearly wrong. Answer C states that the algorithm makes an independence assumption. The Viterbi algorithm does not make any assumptions to independence, it is applied to an HMM. Therefore Answer C is inaccurate. Answer D is also not exact, since the HMM models does not capture an existing independence, but simply makes the (wrong) assumption that the independence exists.

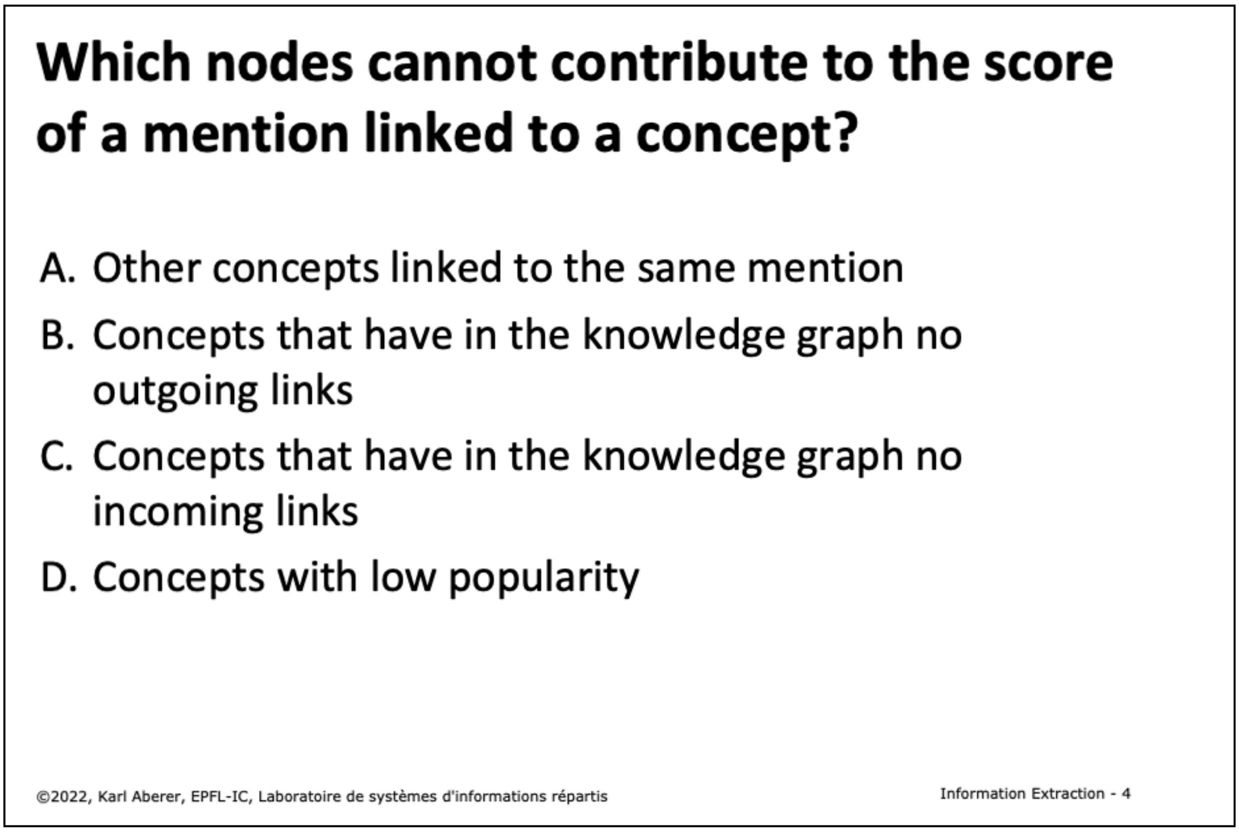


Which is false?

A. Entity disambiguation addresses the problem of synonyms
B. Named entity recognition addresses the problem of synonyms
C. Entity disambiguation addresses the problem of entity classification
D. Named entity recognition addresses the problem of entity classification

Answer B

Names entity recognition is a local method that extracts different entities from text independently. In standard NER the fact that different entities might share the names is not taken into account. However, NER performs classification into different entity types.



Which nodes cannot contribute to the score of a mention linked to a concept?

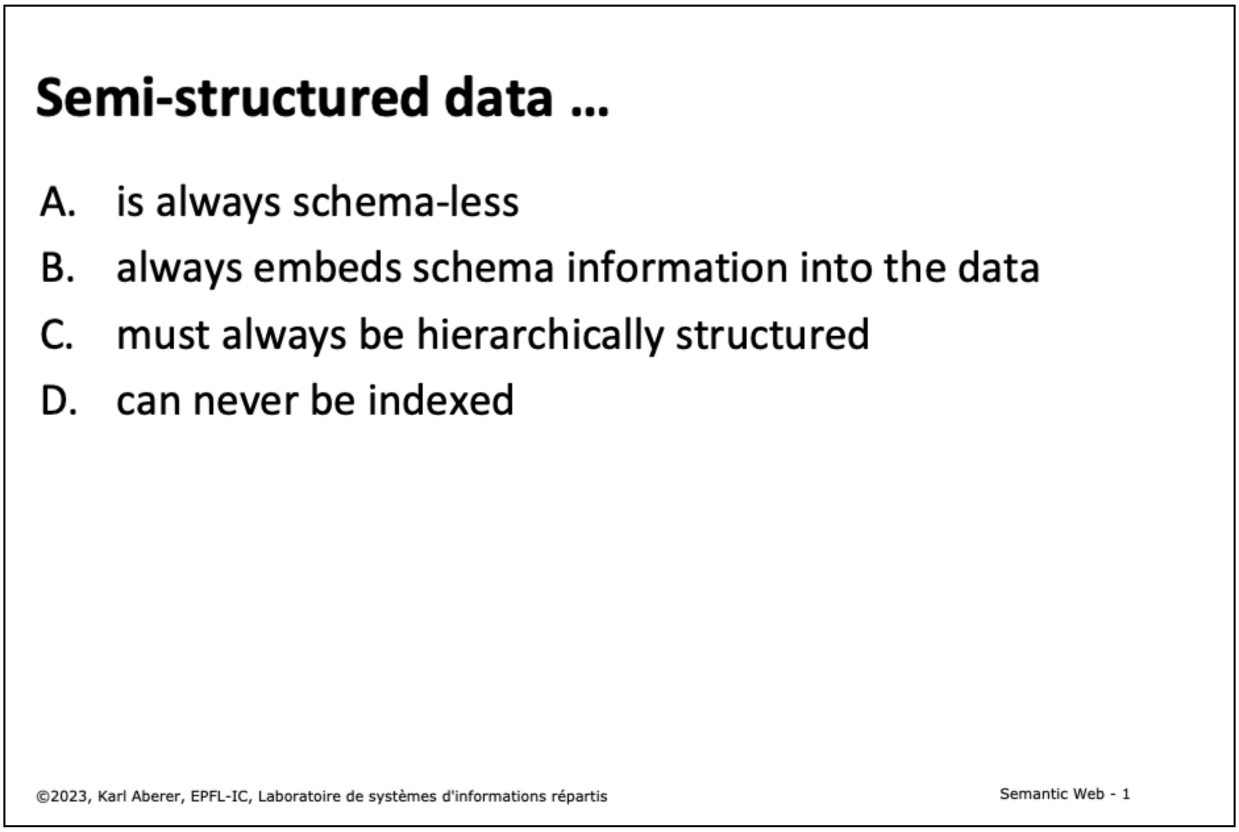
A. Other concepts linked to the same mention
B. Concepts that have in the knowledge graph no outgoing links
C. Concepts that have in the knowledge graph no incoming links
D. Concepts with low popularity

Information Extraction - 4

Answer A

Concepts that have been linked to the same mention are competing for the interpretation. Therefore it is not meaningful to take them into account when trying to support an alternative interpretation.

One might wonder why Answer B and C are correct. A node with no outgoing link can still be the end of a path when computing PPR. Therefore it will contribute to the ranking. A node with no incoming link can on the other hand be the start of a path when comuting PPR and therefore contributes to the scoring.



Semi-structured data …

A. is always schema-less

B. always embeds schema information into the data

C. must always be hierarchically structured

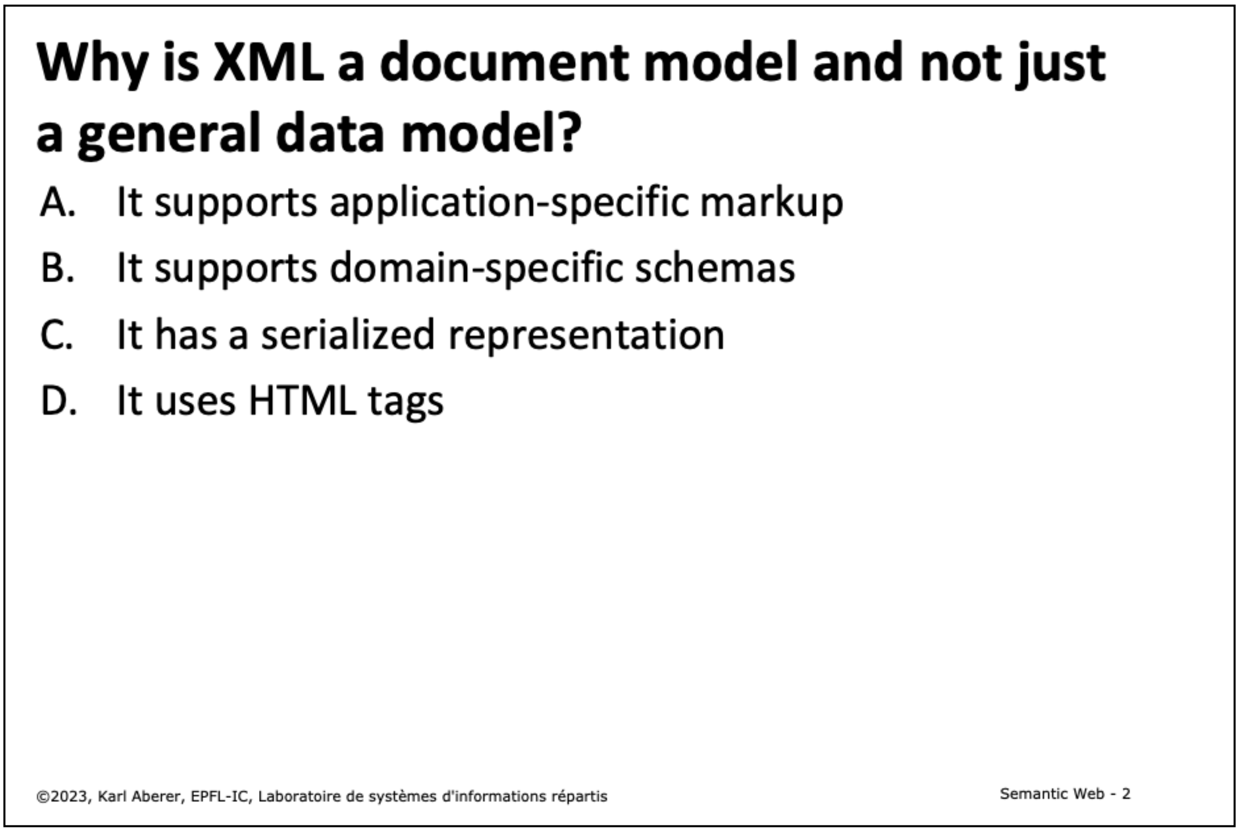
D. can never be indexed

Answer B

Answer A: there exist frameworks for providing schemas for semi-structured data, such as XML Schema, or RDF schema we will introduce.

Answer C: hierarchical structure is very common for semi-structured data, but not necessary, as for example for email formats.

Answer D: Indexing of semi-structured data is possible, but cannot rely on a fixed data structure. Therefore the approaches are different from the ones used typical in databases with structure schemas.



Why is XML a document model and not just a general data model?

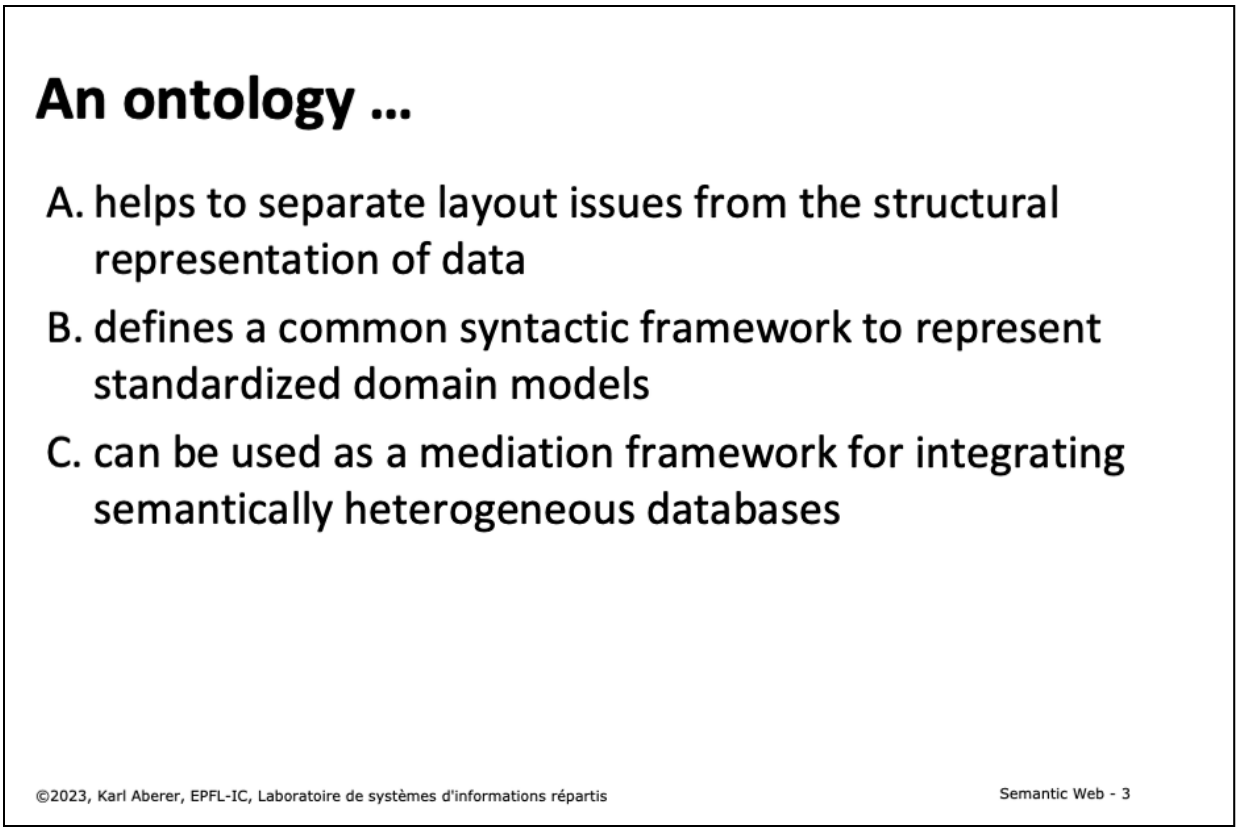
A. It supports application-specific markup

B. It supports domain-specific schemas

C. It has a serialized representation

D. It uses HTML tags

Answer C: the notion of “document” is related to the fact that the data is represented as a string. Statements A and B are also correct for XML, but do not justify the notion of document. More accurately, we are referring to text documents here.



An ontology …

A. helps to separate layout issues from the structural representation of data

B. defines a common syntactic framework to represent standardized domain models

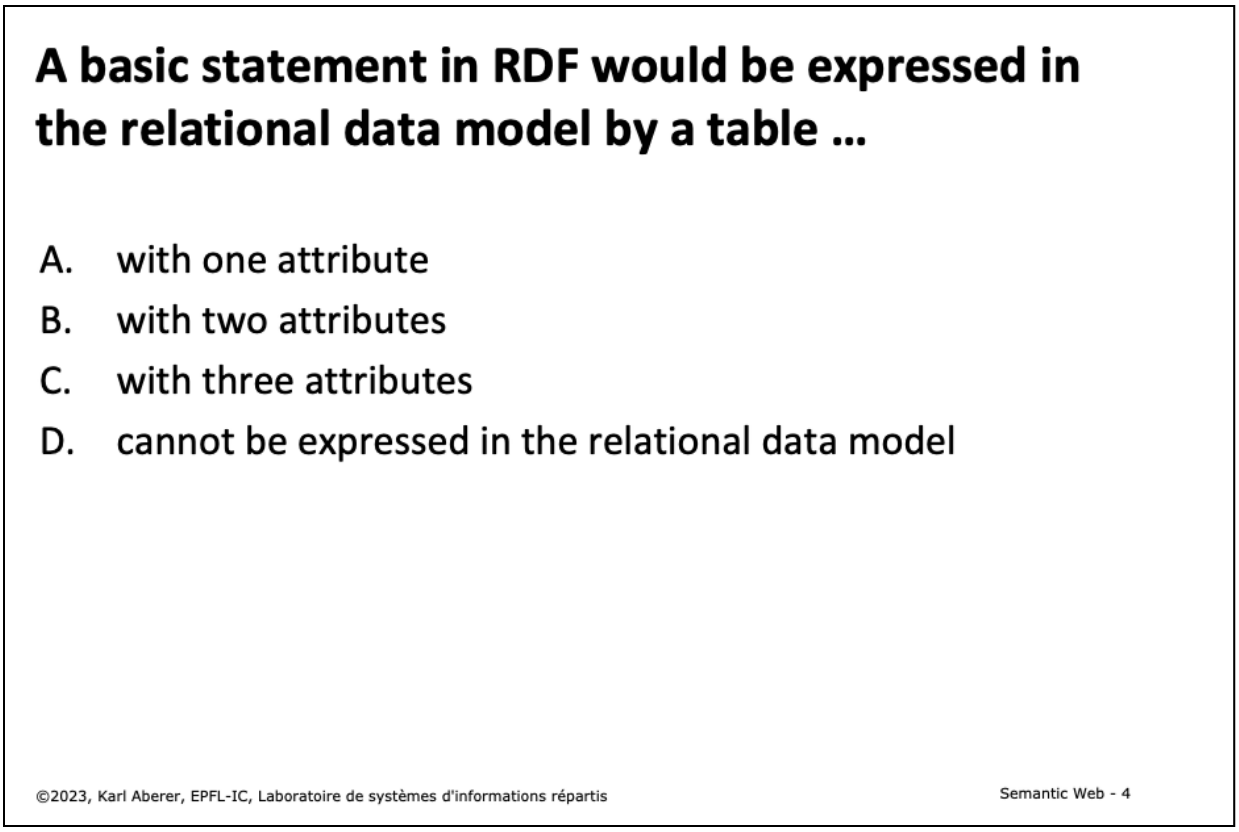
C. can be used as a mediation framework for integrating semantically heterogeneous databases

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Semantic Web - 3

Answer C

Answer A refers to the role of structured document models, such as XML. An ontology is a priori also independent of its specific syntactic representation. It can be used to integrate data from semantically heterogeneous resources.



A basic statement in RDF would be expressed in the relational data model by a table …

A. with one attribute

B. with two attributes

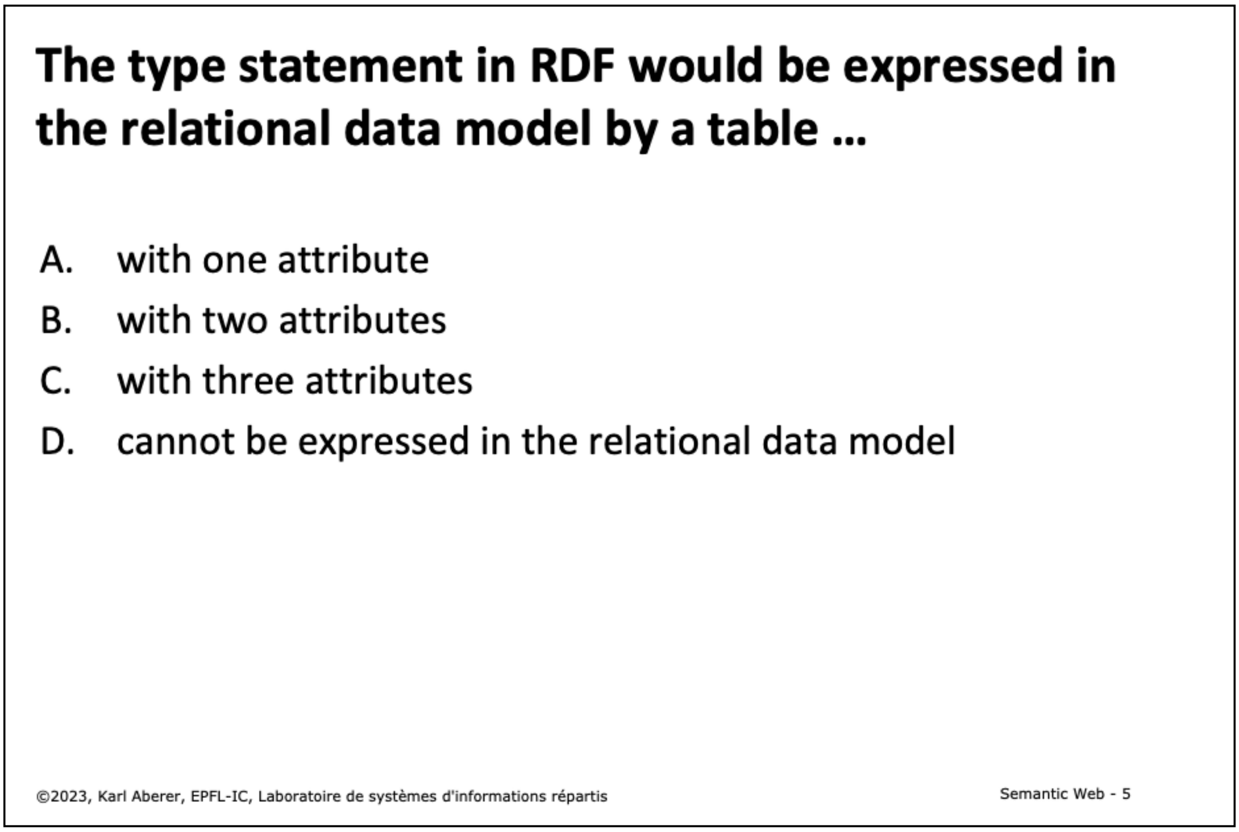
C. with three attributes

D. cannot be expressed in the relational data model

Semantic Web - 4

Answer B

The canonical representation of an RDF statement in a relational table would be by a table that represents the predictate of the statement and two attributes that are used to represent the subject and object. For example, the statement X isauthor Y would be represented as tuple is-author(X, Y). One might argue that also the predicate might be represented as attribute. This would be resulting in a generic representation with single table used to represent all RDF statements. This is not the standard approach of data modeling used in relational data models.



The type statement in RDF would be expressed in the relational data model by a table …

A. with one attribute

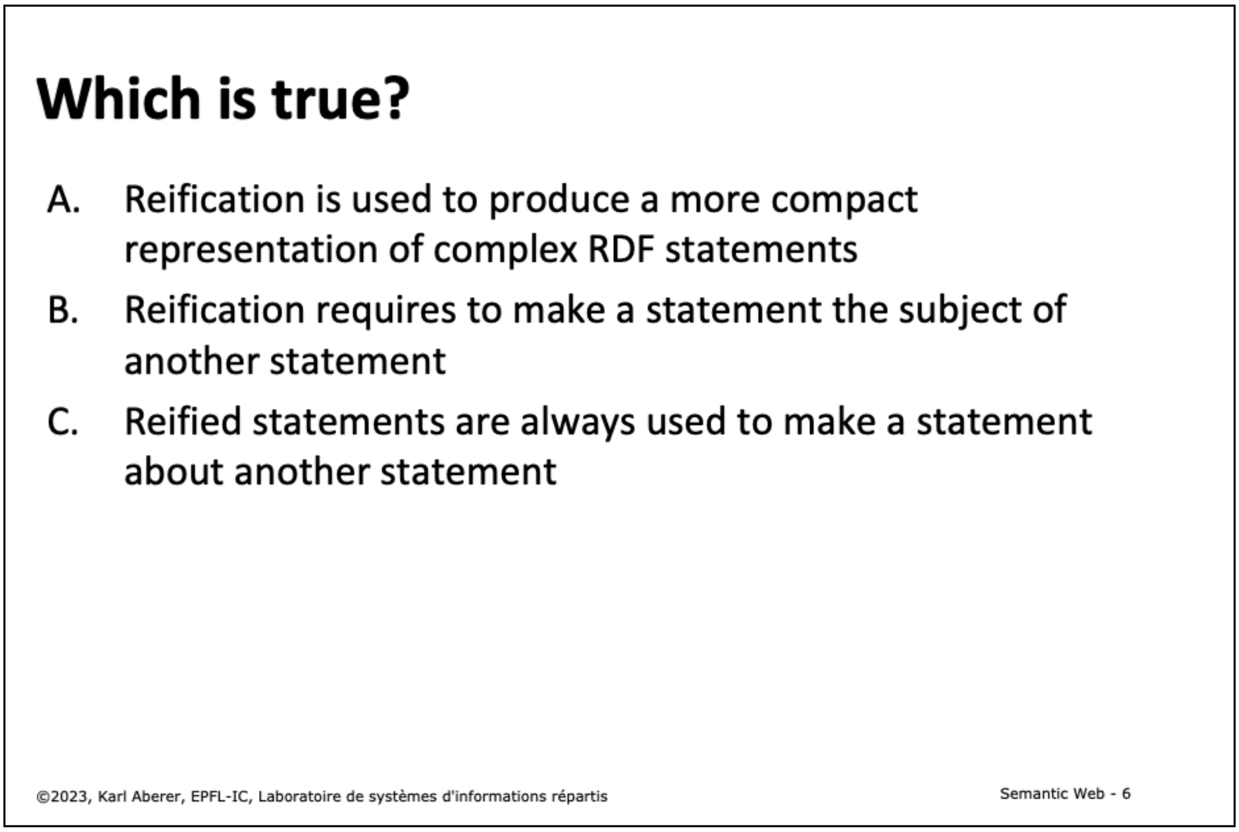
B. with two attributes

C. with three attributes

D. cannot be expressed in the relational data model

Answer A

The type relation expresses membership to a sepcific category of elements in a domain. The canonical way to represent this in the relational model is to use a table that lists the elements of the category. For example, if we have elements of type country, and statements of the form X is-a country, we would represent this by tuples of the form country(X). One might argue that one could list the type also as an attribute in a binary table. However, then this corresponds rather to a specific property than a type information. For example, if we have elements of type female and male, we would have a table gender(X, F/M), but we would consider the gender not as type information.



Which is true?

A. Reification is used to produce a more compact representation of complex RDF statements

B. Reification requires to make a statement the subject of another statement

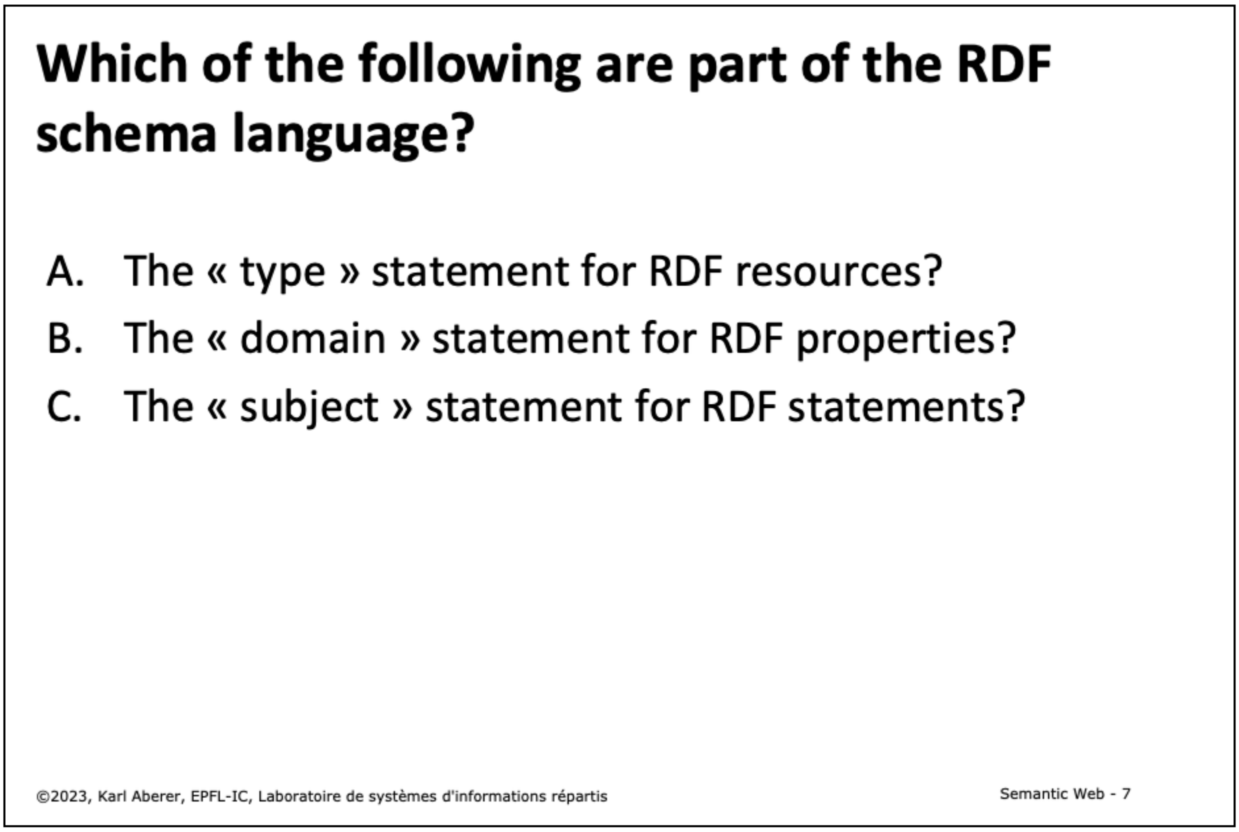
C. Reified statements are always used to make a statement about another statement

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Semantic Web - 6

Answer C

Answer A is obviously the opposite what happens. As for Answer B, a statement about which we make another statement can also be the object of the statement. For example, author X produced the statement S.



Which of the following are part of the RDF schema language?

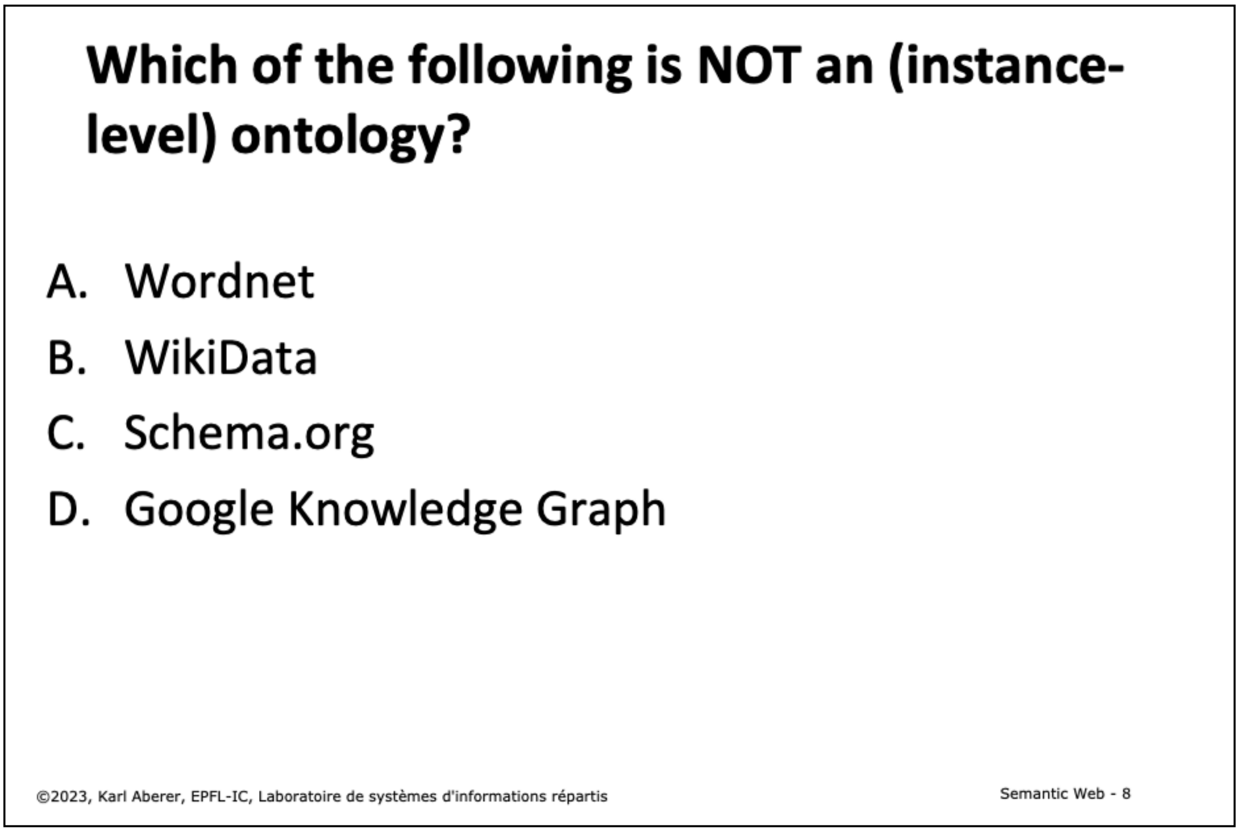
A. The « type » statement for RDF resources?

B. The « domain » statement for RDF properties?

C. The « subject » statement for RDF statements?

Answer B

You have noticed that the namespace for the type and subject statements is RDF. This makes sense since these two types of statements involve instances that are not part of an RDF schema, wheras the domain statement involves exclusively schema instances.



Which of the following is NOT an (instance-level) ontology?

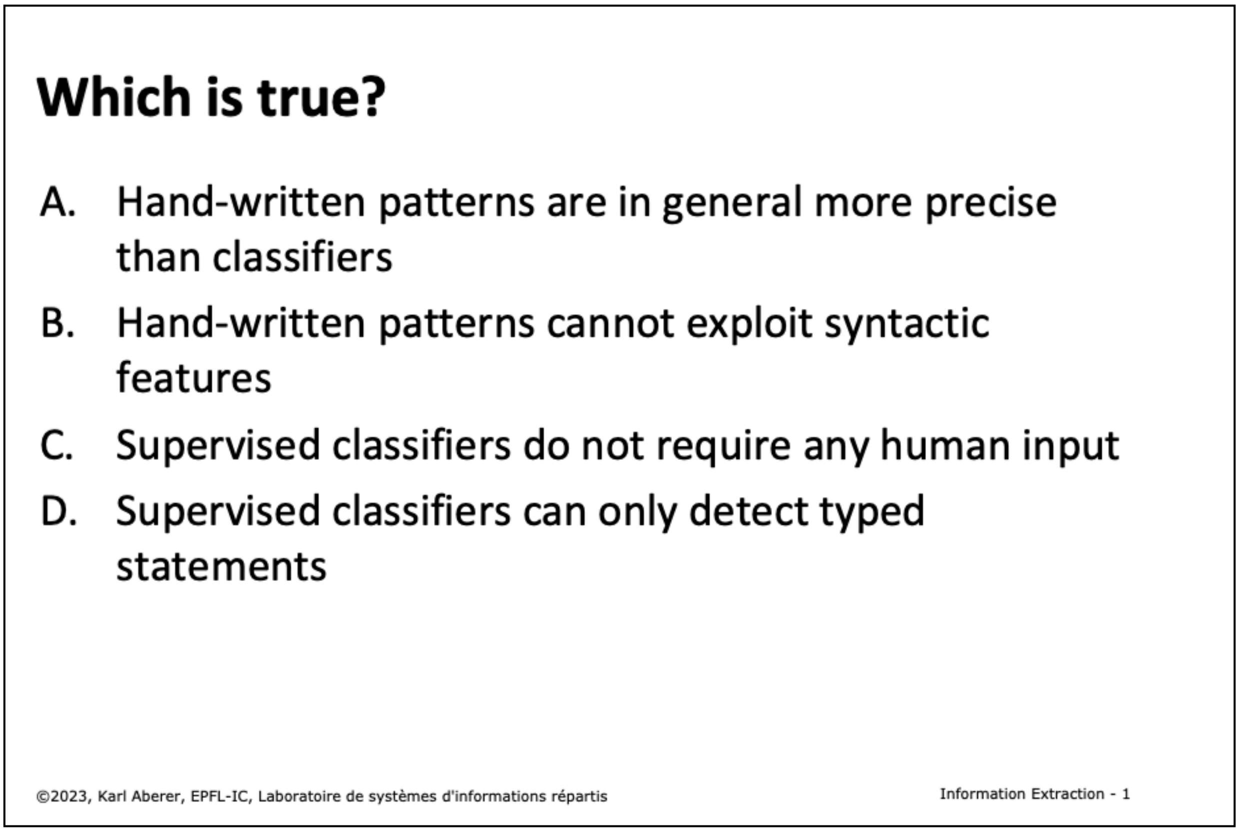
A. Wordnet
B. WikiData
C. Schema.org
D. Google Knowledge Graph

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Semantic Web - 8

Answer C

Schema.org provides a schema without any instances of facts, whereas the three other resources provide sets of facts.



Which is true?

A. Hand-written patterns are in general more precise than classifiers

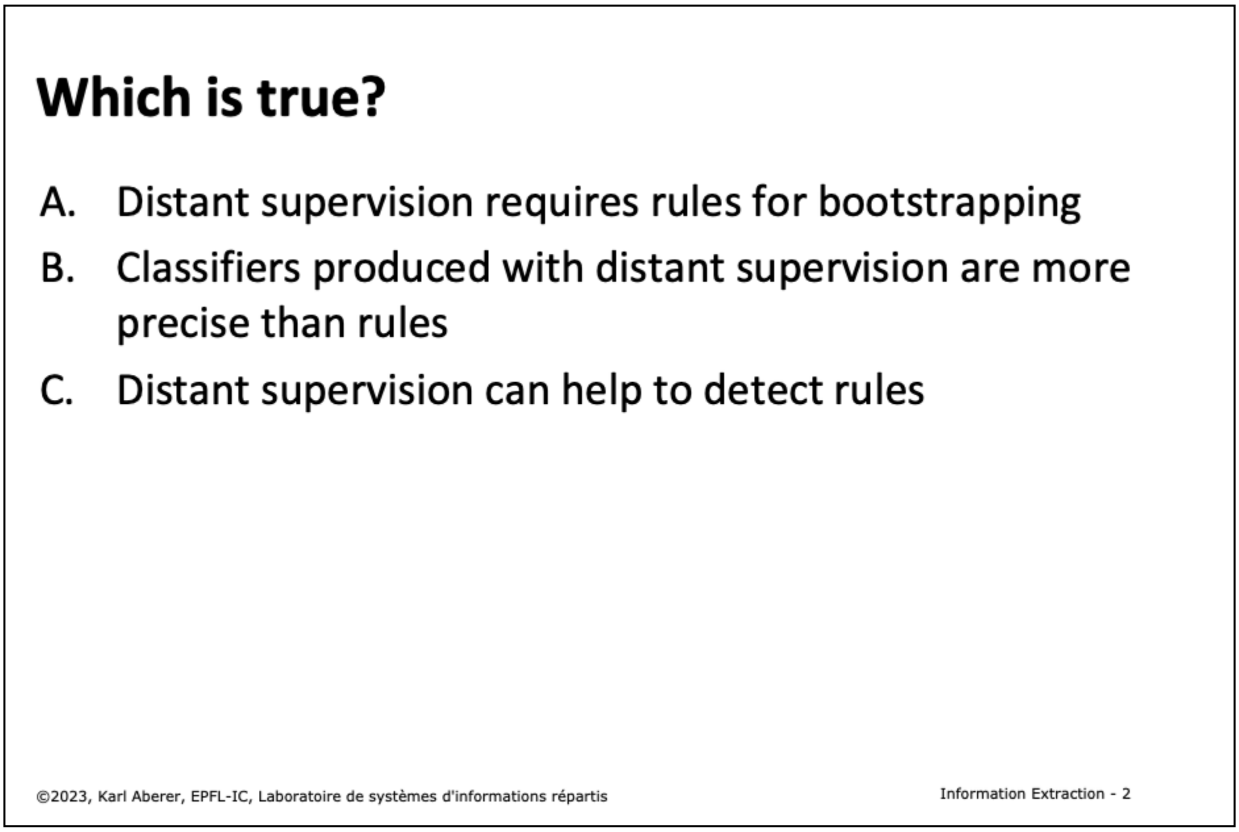
B. Hand-written patterns cannot exploit syntactic features

C. Supervised classifiers do not require any human input

D. Supervised classifiers can only detect typed statements

Answer A

Answer B is exactly the opposite of what is the case, that patterns exploit syntactice features. Classifiers requires human input for labelling. They also can be used to detet untyped statements.



Which is true?

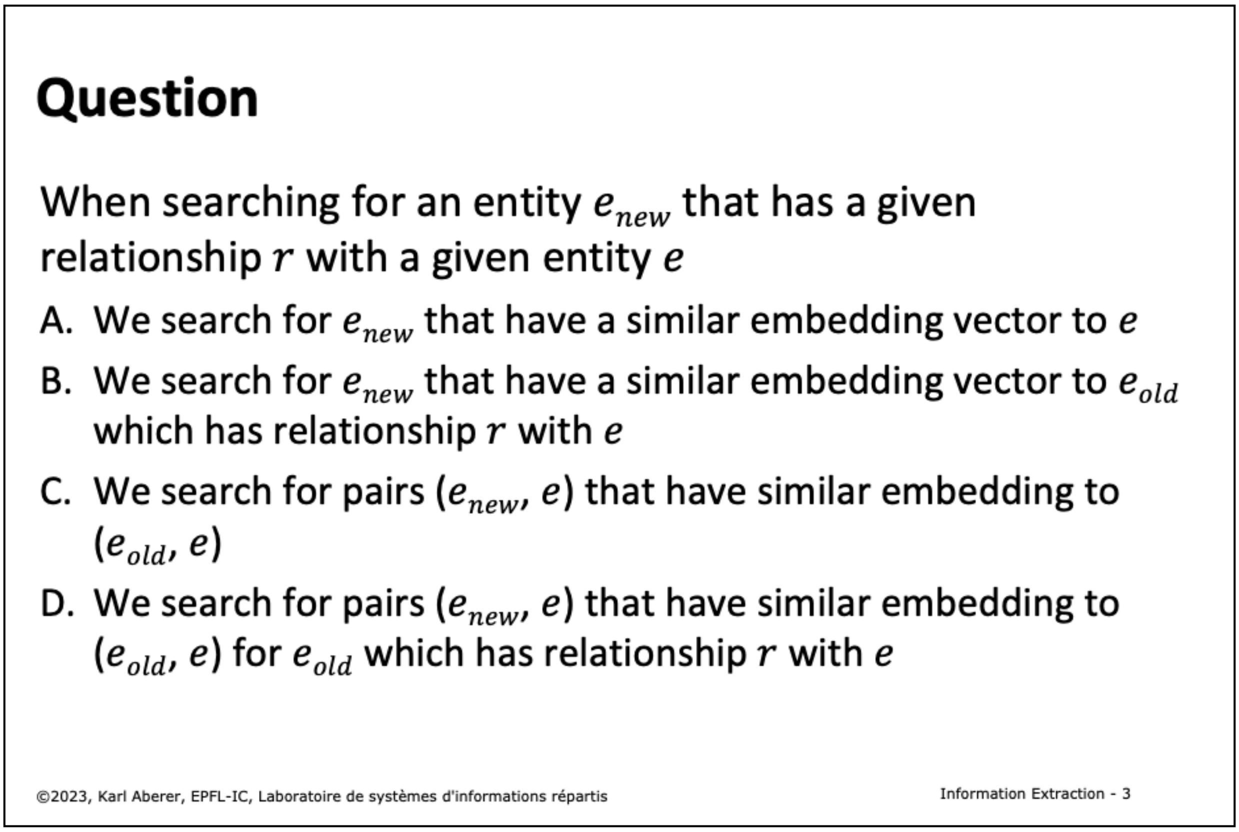
A. Distant supervision requires rules for bootstrapping

B. Classifiers produced with distant supervision are more precise than rules

C. Distant supervision can help to detect rules

Answer C

Distant supervision does not require and rules, as opposed to bootstrapping. Classifiers from distant supervision tend to be less precise than hand-written rules, though in some cases they achieve comparable performance. Using complex features, distant supervision effectively detects new rules.



Question

When searching for an entity e_{new} that has a given relationship r with a given entity e

A. We search for e_{new} that have a similar embedding vector to e

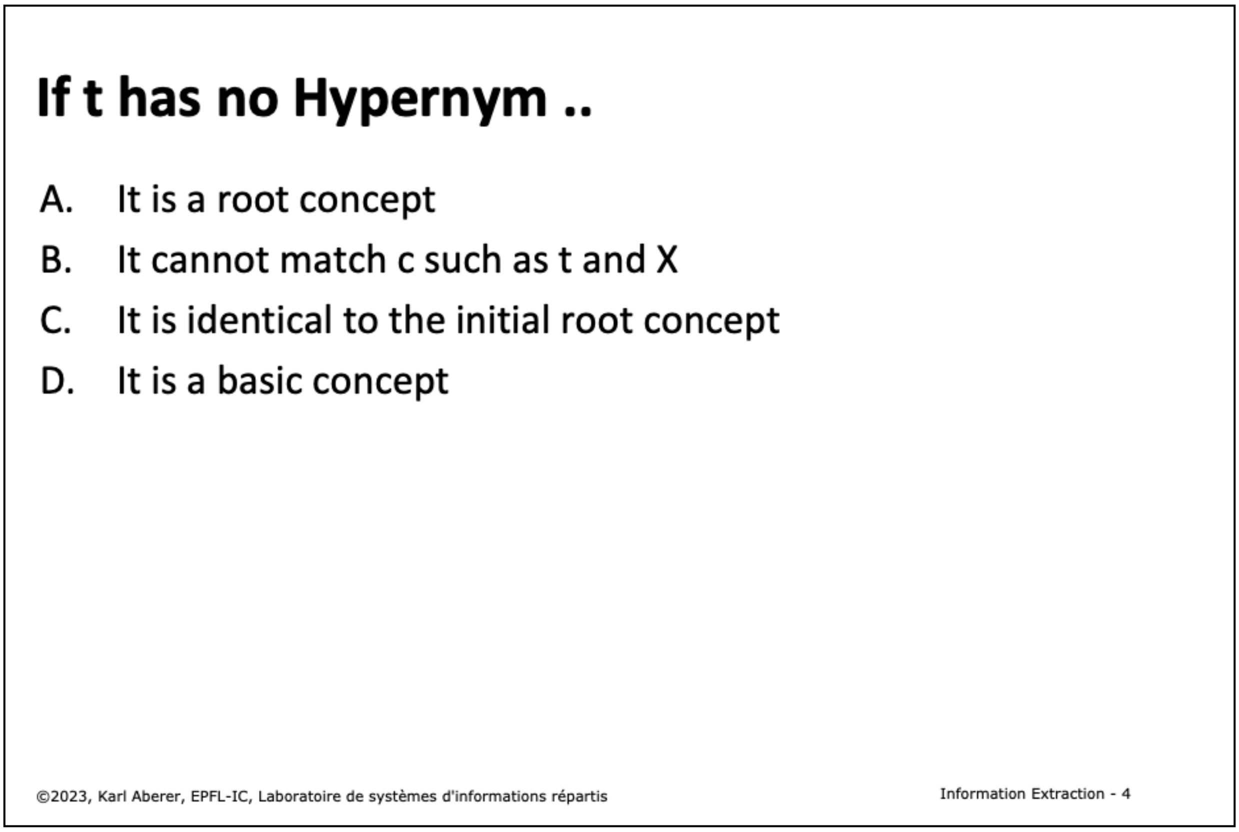
B. We search for e_{new} that have a similar embedding vector to e_{old} which has relationship r with e

C. We search for pairs (e_{new}, e) that have similar embedding to (e_{old}, e)

D. We search for pairs (e_{new}, e) that have similar embedding to (e_{old}, e) for e_{old} which has relationship r with e

Answer C

In the model for relation embedding individual entities have no embedding, therefore Answers A and B are not applicable. As for Answer D, it is not required that the existing embedding pairs (e_{old}, e) have a confirmed relationship. However, we could also search for embeddings of relations r that are similar to the embedding (e_{new}, e).



If t has no Hypernym ..

A. It is a root concept

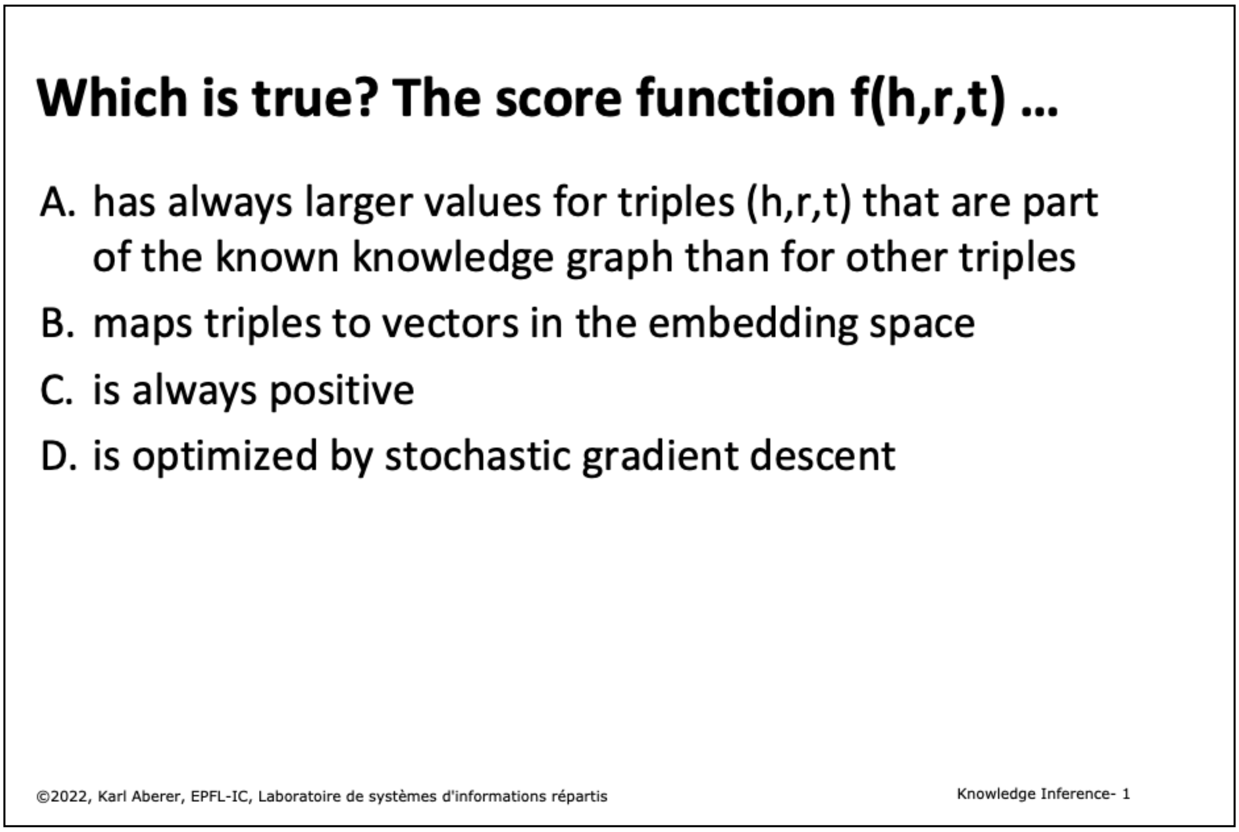
B. It cannot match c such as t and X

C. It is identical to the initial root concept

D. It is a basic concept

Answer A

Hypernym means that the concept has no more general concept, therefore it is a root concept. It cannot be part of the pattern from Answer B, since c would be a hypernym of t. In the search for hypernyms different root concepts can be found that have no relation among each other, which excludes Answer C.



Which is true? The score function f(h,r,t) …

A. has always larger values for triples (h,r,t) that are part of the known knowledge graph than for other triples

B. maps triples to vectors in the embedding space

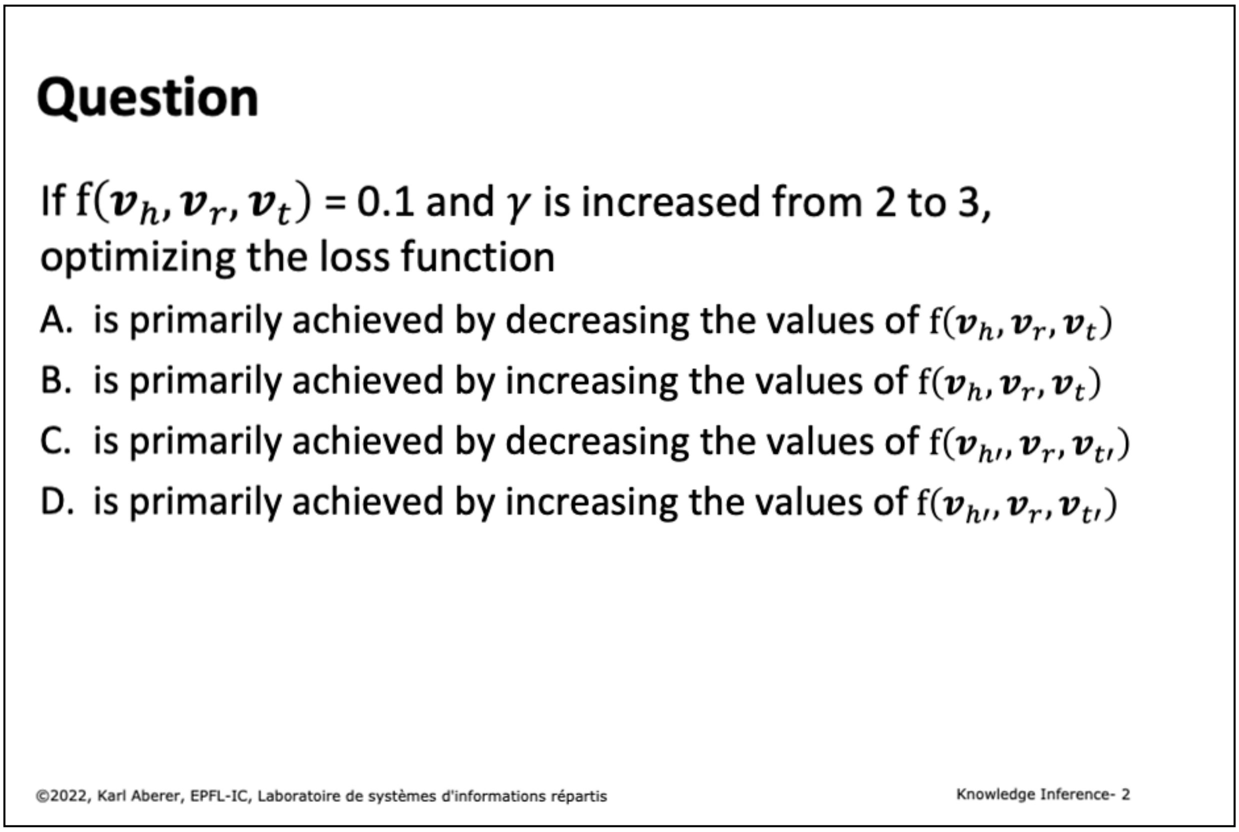
C. is always positive

D. is optimized by stochastic gradient descent

Answer C

The score function is a dissimilarity function that goes to zero when the facts are correct.

Answer A is exactly the opposite of the behavior of the score functions. As for Answer B, the score function maps tripes into real numbers, not vectors. And finally SGD optimizes the loss function not the score function.



Question

If $\mathrm{f}(\boldsymbol{v}_h, \boldsymbol{v}_r, \boldsymbol{v}_t)$ = 0.1 and γ is increased from 2 to 3, optimizing the loss function

A. is primarily achieved by decreasing the values of $\mathrm{f}(\boldsymbol{v}_h, \boldsymbol{v}_r, \boldsymbol{v}_t)$

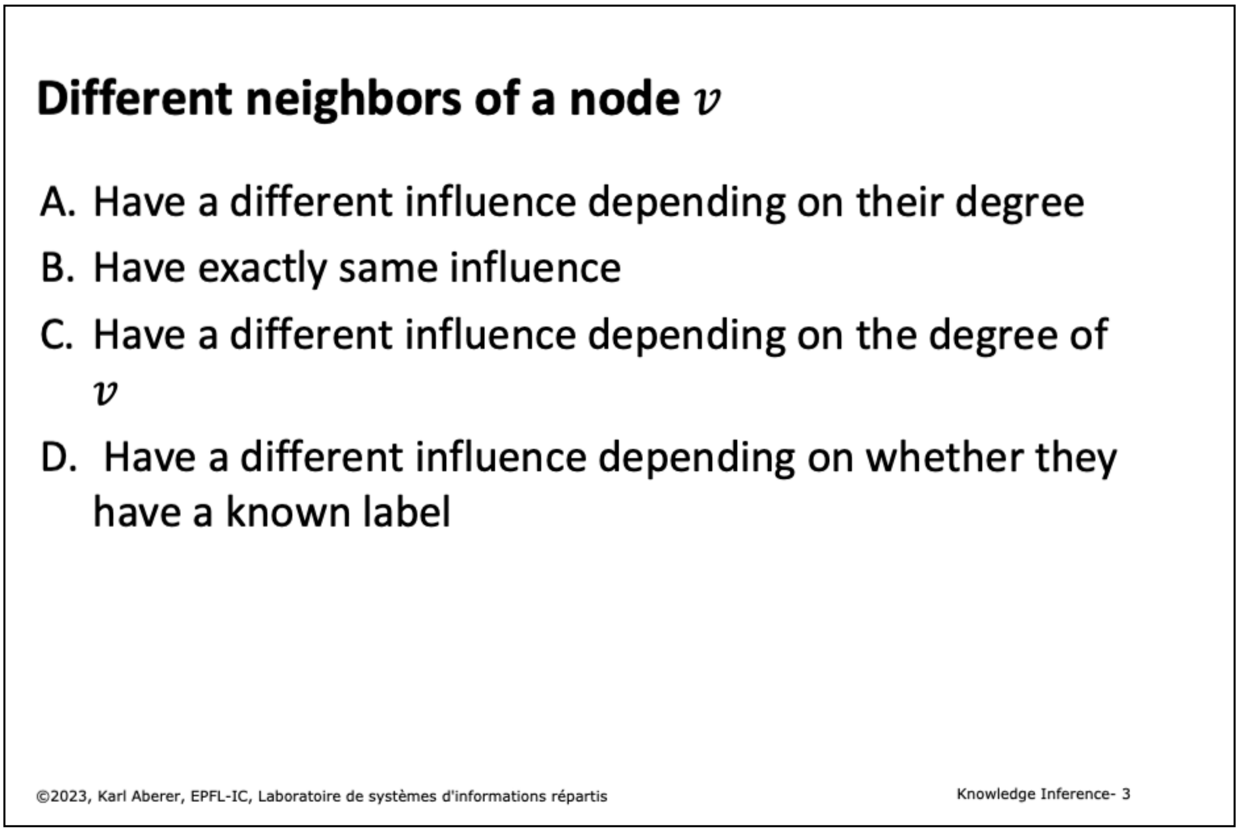
B. is primarily achieved by increasing the values of $\mathrm{f}(\boldsymbol{v}_h, \boldsymbol{v}_r, \boldsymbol{v}_t)$

C. is primarily achieved by decreasing the values of $\mathrm{f}(\boldsymbol{v}_{h'}, \boldsymbol{v}_r, \boldsymbol{v}_{t'})$

D. is primarily achieved by increasing the values of $\mathrm{f}(\boldsymbol{v}_{h'}, \boldsymbol{v}_r, \boldsymbol{v}_{t'})$

Answer D

Since the score function for $\mathrm{f}(\boldsymbol{v}_h, \boldsymbol{v}_r, \boldsymbol{v}_t)$ us already almost zero, the increase in margin can only be achieved by increasing the scores for the implausible triples, $\mathrm{f}(\boldsymbol{v}_{h'}, \boldsymbol{v}_r, \boldsymbol{v}_{t'})$.



Different neighbors of a node v

A. Have a different influence depending on their degree

B. Have exactly same influence

C. Have a different influence depending on the degree of v

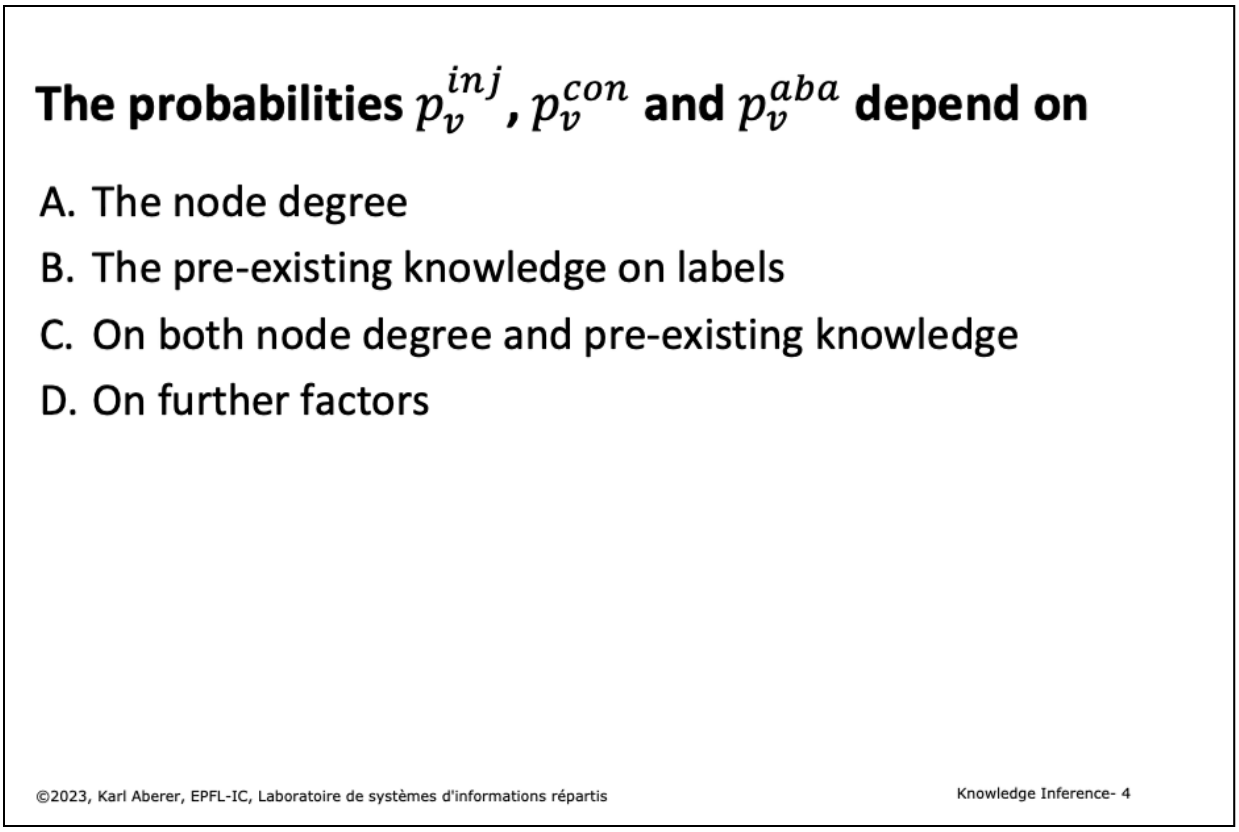
D. Have a different influence depending on whether they have a known label

Answer C is incorrect, since the degree of a node v has no influence of how it is updated by its neighbors.

For a similar reason also answer D is incorrect. The fact that a node has a known label does not influence how it is updated by the neighbors.

Answer B is not correct since neighbors have different influences, e.g. depending on their node degrees.

Answer A is correct, as the degree of neighbors determines their influence.



The probabilities p_v^{inj}, p_v^{con} and p_v^{aba} depend on

A. The node degree

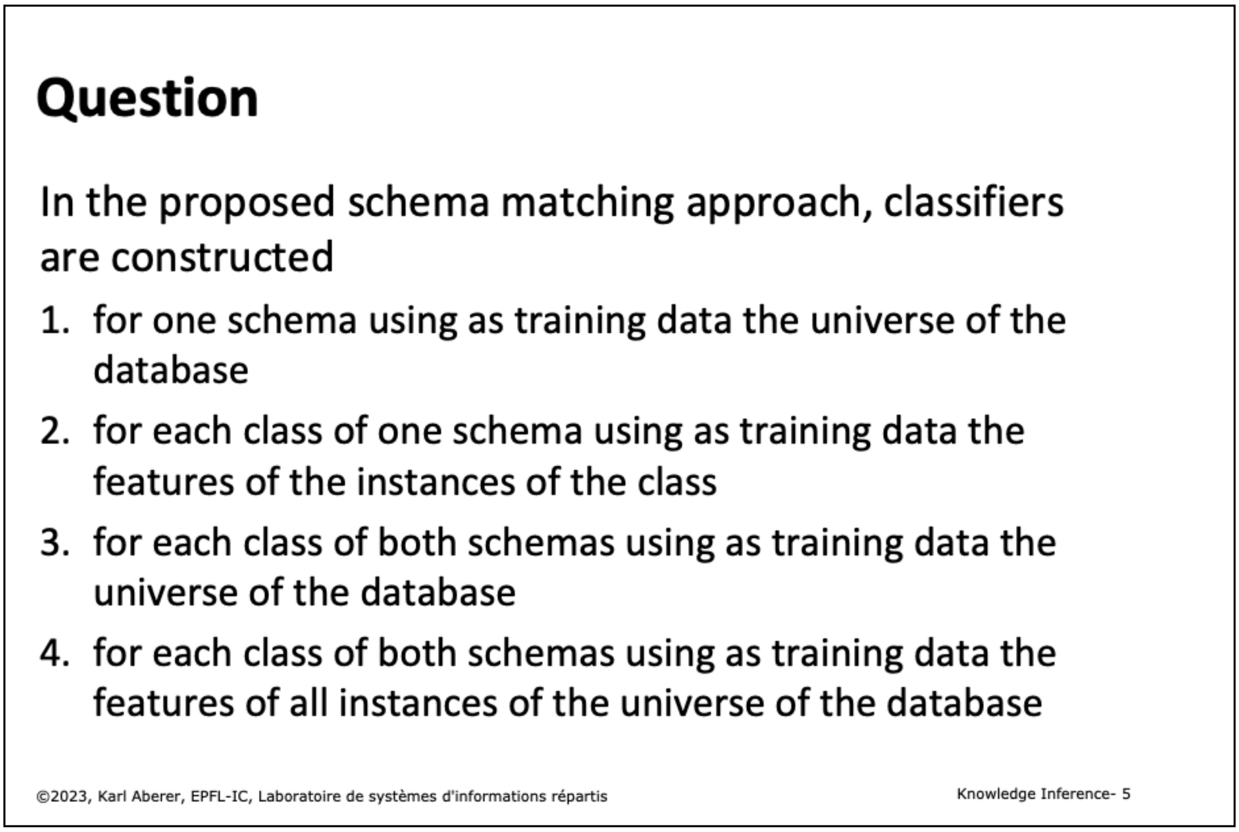
B. The pre-existing knowledge on labels

C. On both node degree and pre-existing knowledge

D. On further factors

Answer C

The probabilities take clearly into account the node degrees. For each node they also depend on whether a pre-existing label exist or not, thus depend on the pre-existing knowledge.



Question

In the proposed schema matching approach, classifiers are constructed

1. for one schema using as training data the universe of the database
2. for each class of one schema using as training data the features of the instances of the class
3. for each class of both schemas using as training data the universe of the database
4. for each class of both schemas using as training data the features of all instances of the universe of the database

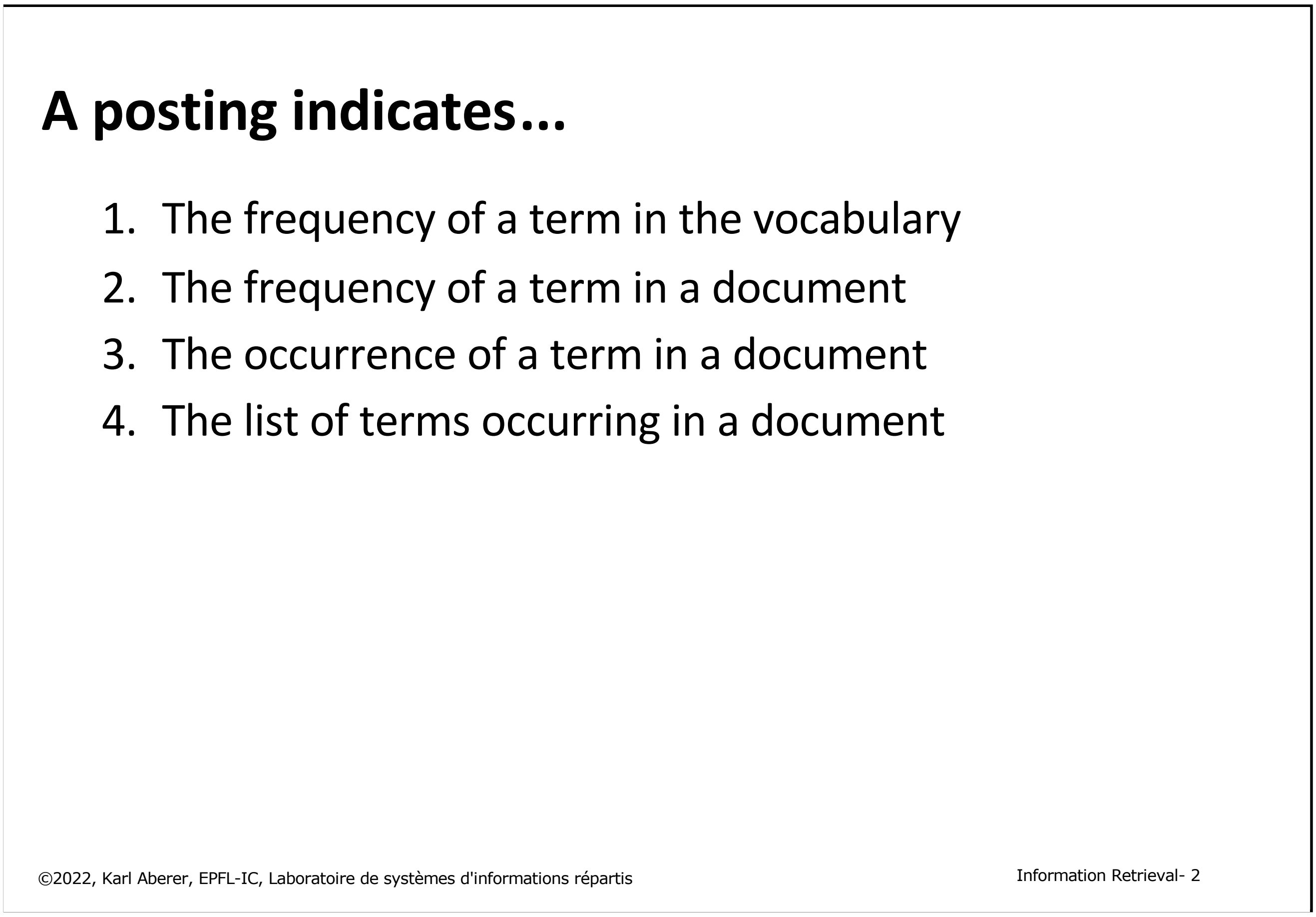
Answer 4

The classifiers are trained using the features of instances of the universe (this excludes answers 1 and 3 already).

The classifiers are trained per class, for both schemas, using as training data the features of the whole universe, which serve as positive and negative examples of the members of the class.

Answers to quizzes

5.1 INDEXING FOR INFORMATION RETRIEVAL

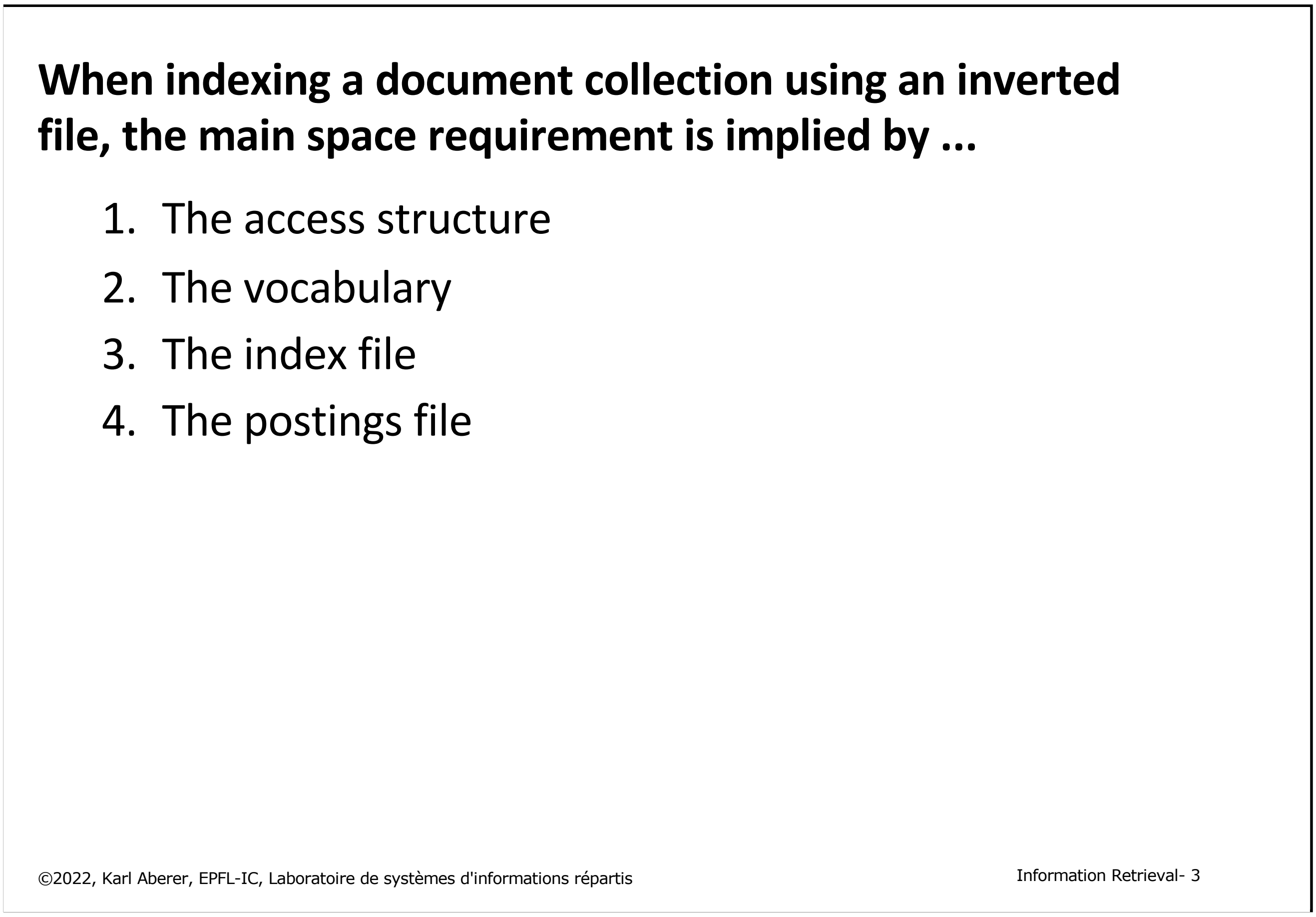


A posting indicates...

1. The frequency of a term in the vocabulary
2. The frequency of a term in a document
3. The occurrence of a term in a document
4. The list of terms occurring in a document

Answer 3

A posting corresponds to the occurrence (or location) of a term in a document. It may contain in its payload the frequency of occurrence of the term, so Answer 2 might be true in special cases, but not in general.

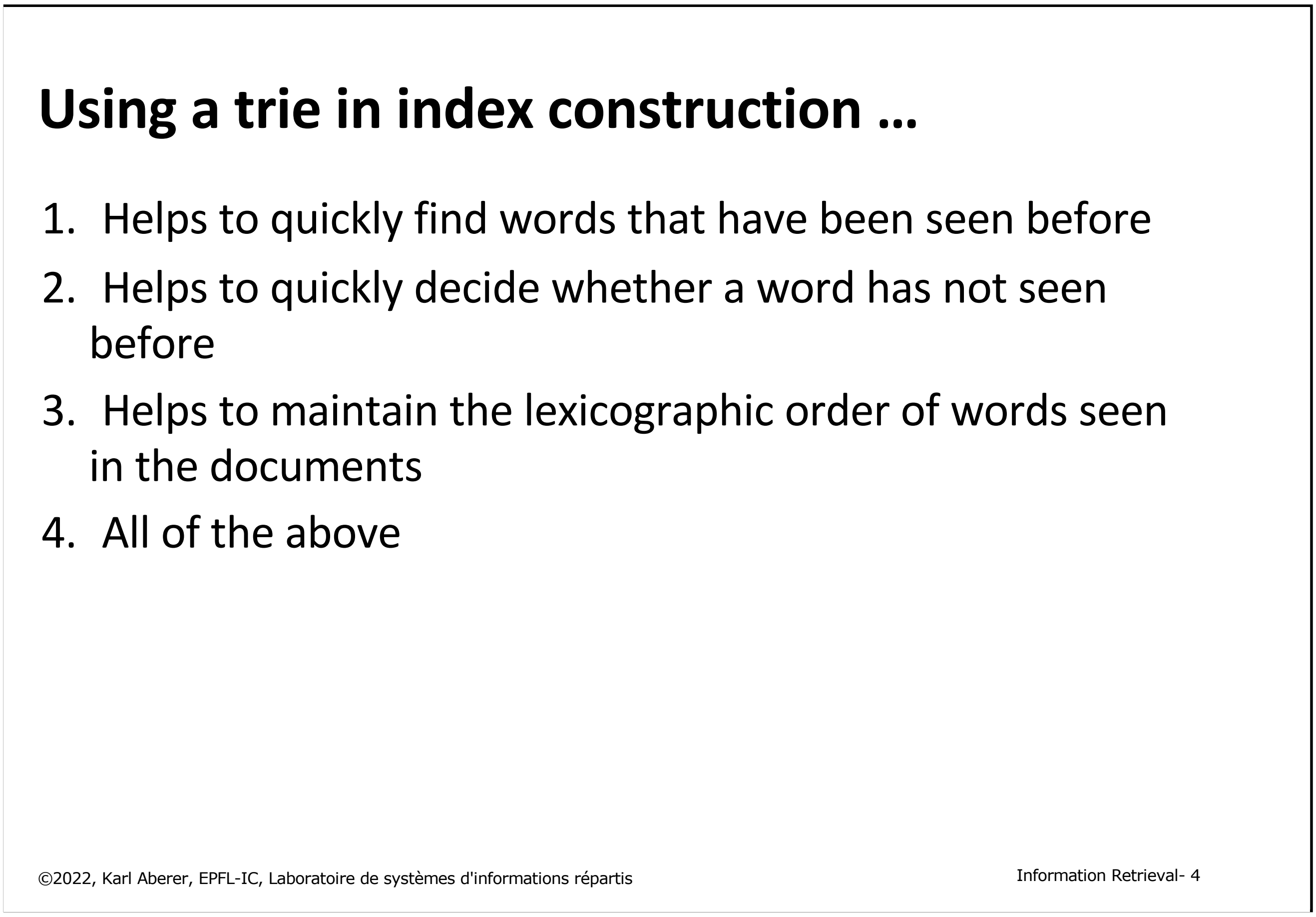


When indexing a document collection using an inverted file, the main space requirement is implied by ...

1. The access structure
2. The vocabulary
3. The index file
4. The postings file

Answer 4

The sizes of the access structure, the vocabulary and the index file all are proportional to the size of the vocabulary which follows for reald world datasets Heap’s law and are therefore of order O(n^0.5). The postings file always has a size of order O(n).



Using a trie in index construction …

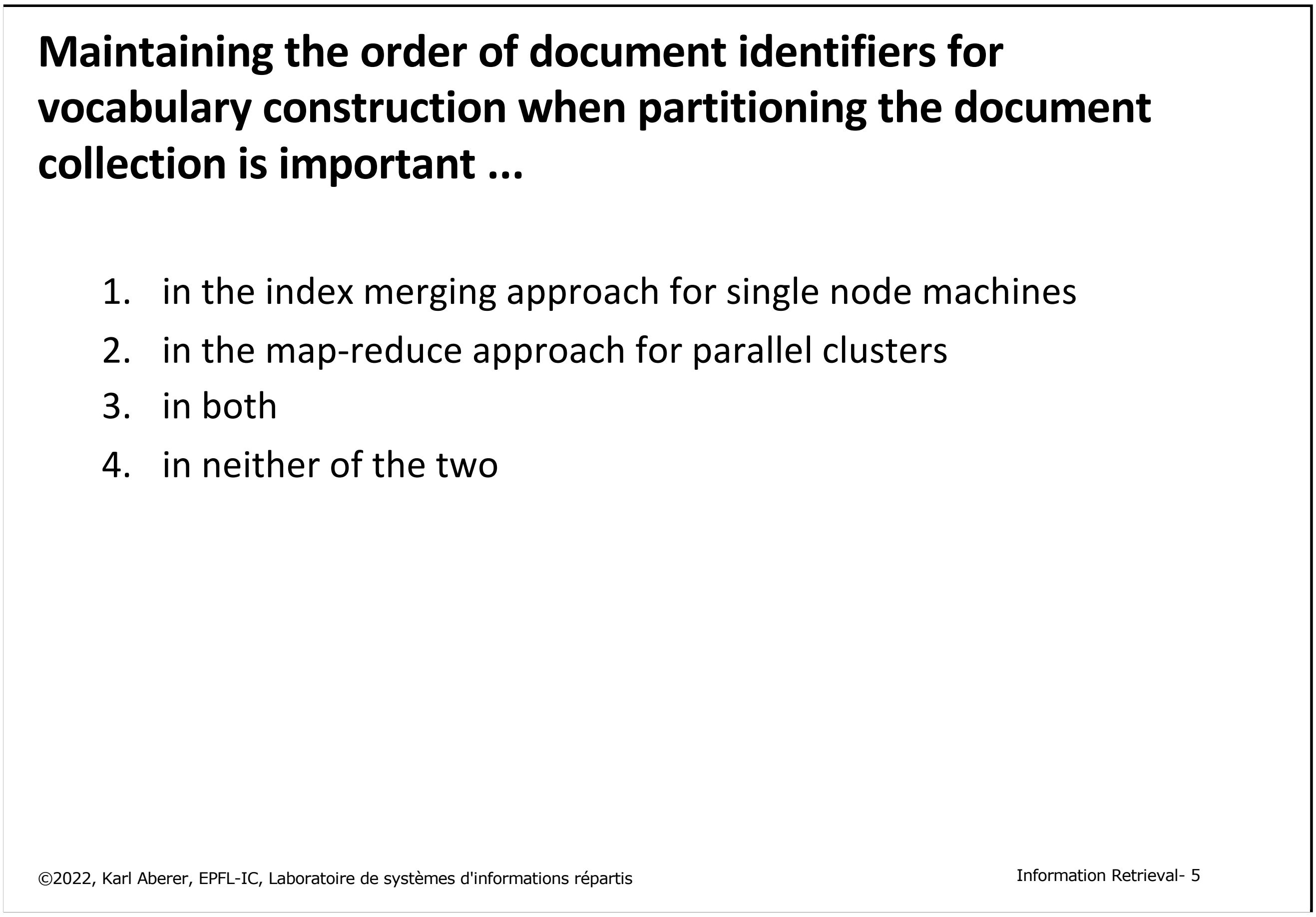
1. Helps to quickly find words that have been seen before
2. Helps to quickly decide whether a word has not seen before
3. Helps to maintain the lexicographic order of words seen in the documents
4. All of the above

©2022, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Information Retrieval- 4

Answer 4

When receiving the next word during the indexing process first the word is searched in the trie constructed so far. So the trie helps to decide whether the word is already known, and if it is known find the associated data. The true can also be used to derive the list of words in. the vocabulary in lexicographical order by traversing the trie. So no sorting of the vocabulary is required.

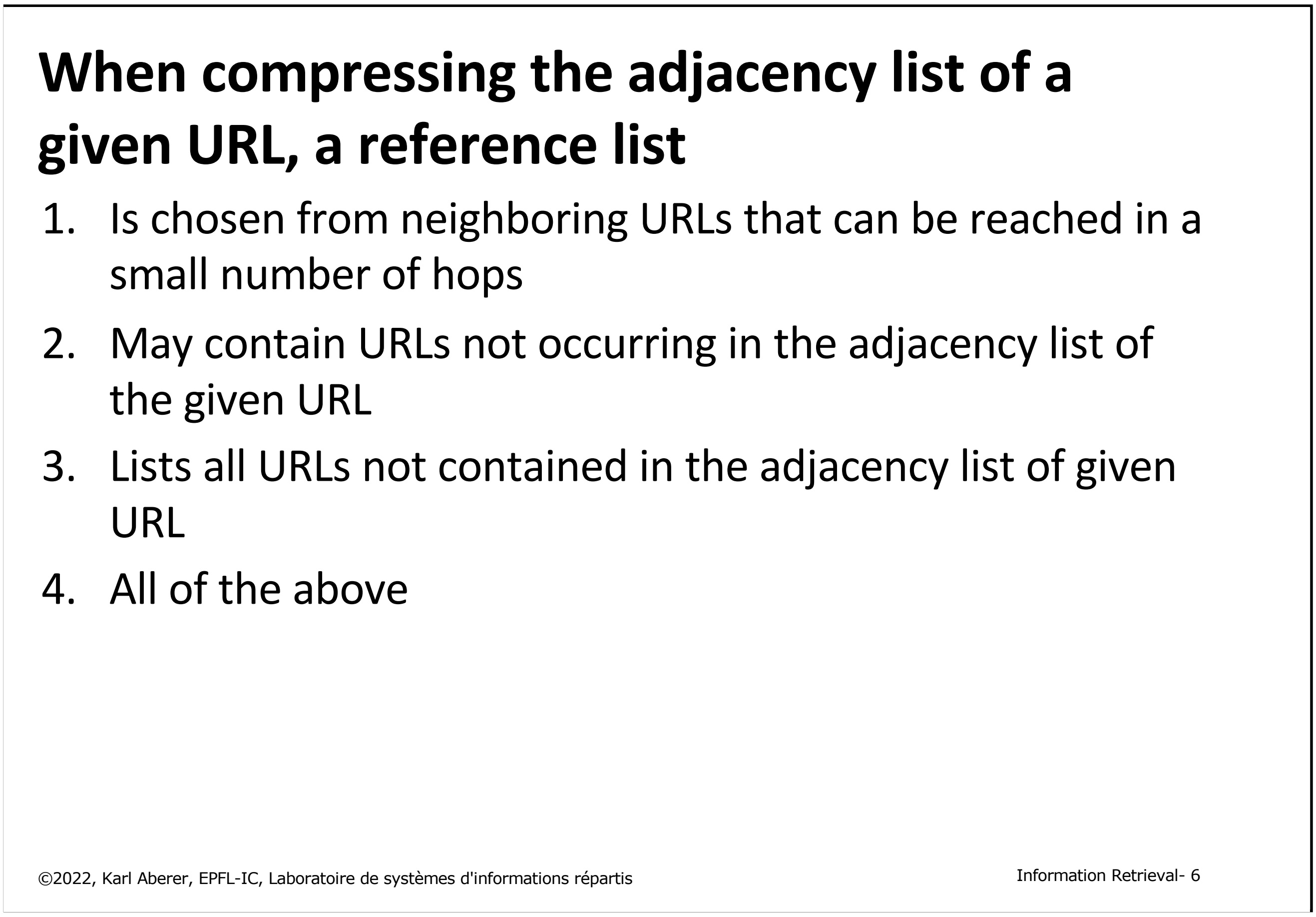


Maintaining the order of document identifiers for vocabulary construction when partitioning the document collection is important ...

1. in the index merging approach for single node machines
2. in the map-reduce approach for parallel clusters
3. in both
4. in neither of the two

Answer 1

In the index merging approach documents are traversed in order and the order is maintained when generating the postings for individual words, which is important for avoiding expensive sorting. In the map-reduce indexing approach postings are sent over the network to the reducer nodes in an arbitrary order, since the infrastructure is controlling the communication of the messages. Therefore, the reducer nodes have to reestablish the order of document identifiers.

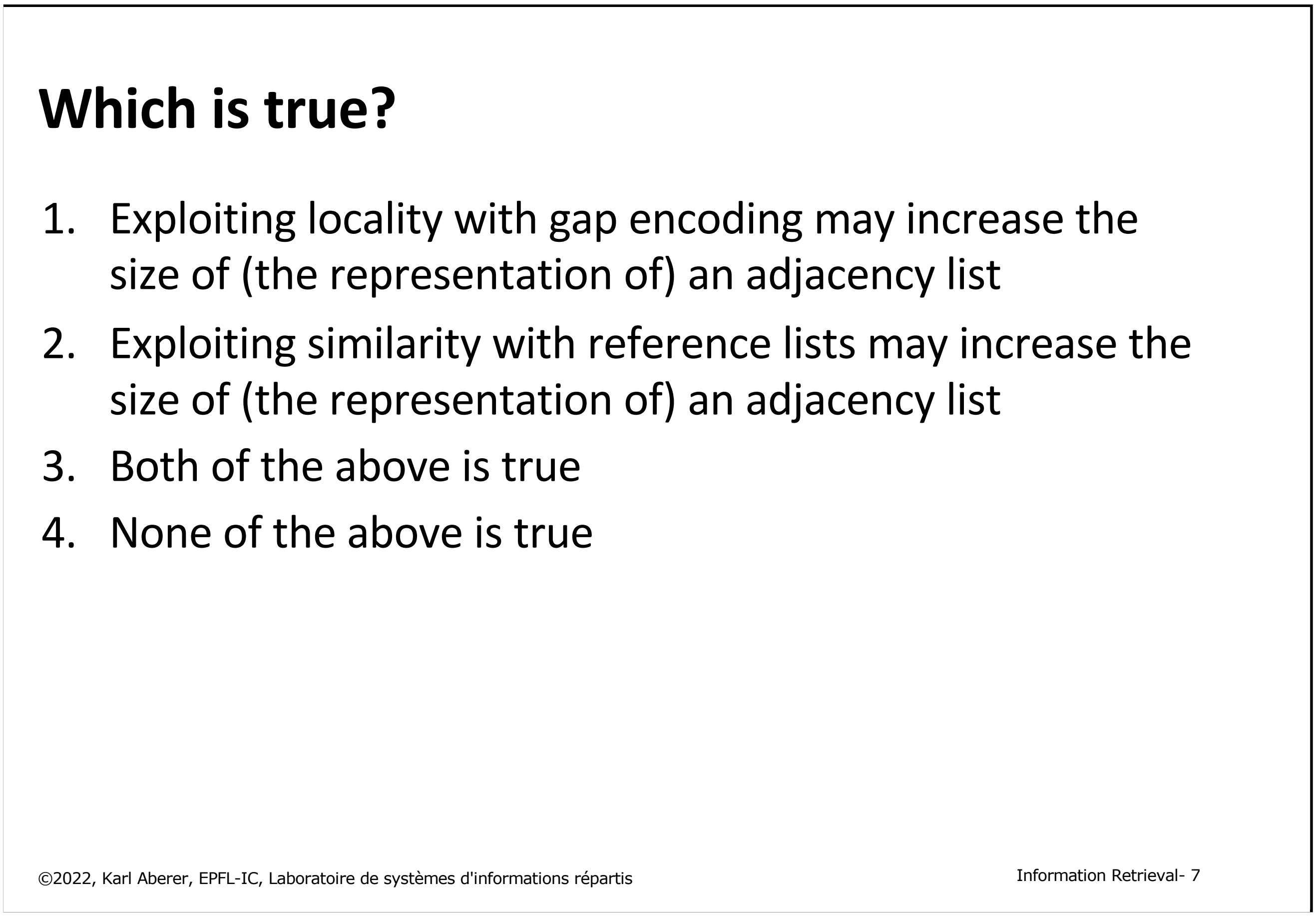


When compressing the adjacency list of a given URL, a reference list

1. Is chosen from neighboring URLs that can be reached in a small number of hops
2. May contain URLs not occurring in the adjacency list of the given URL
3. Lists all URLs not contained in the adjacency list of given URL
4. All of the above

Answer 2

In fact, the reference list can contain URLs that are not contained in the page of the given URL. Answer 1 is wrong since the reference list is chosen from a window in the lexicographical ordering of URLs. Answer 3 does not make sense, as it says that the reference list contains all URLs, except those in the given page.



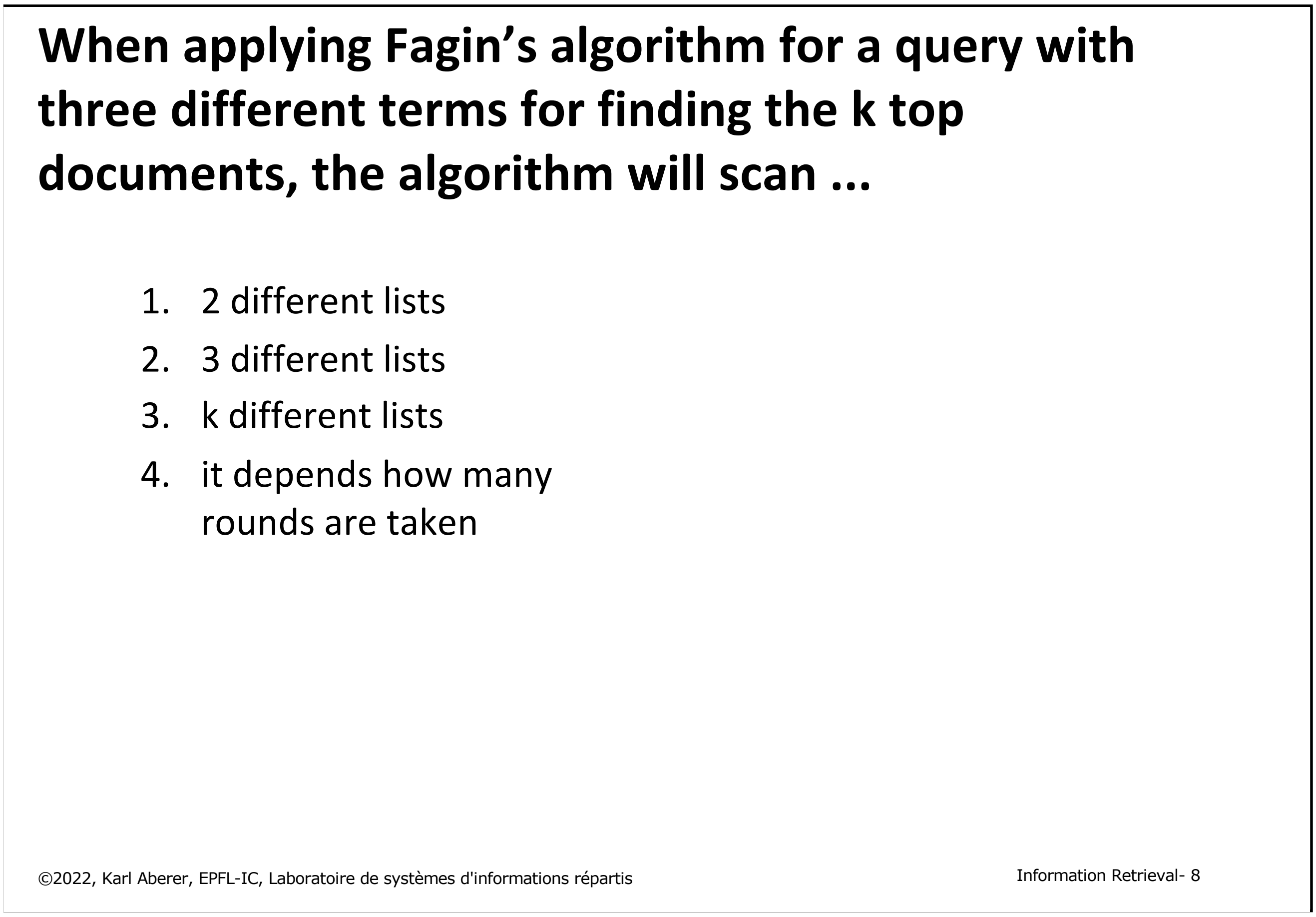
Which is true?

1. Exploiting locality with gap encoding may increase the size of (the representation of) an adjacency list
2. Exploiting similarity with reference lists may increase the size of (the representation of) an adjacency list
3. Both of the above is true
4. None of the above is true

Answer 1

It is possible that with a very unfortuntate distribution of integers the size of the representation might increase with gap encoding (though it is not straightforward to find such a counter-example).

The encoding using a reference list might naturally increase the size of the representation of the adjacency list, if the reference list is poorly chosen. However, in the practical implementation only reference lists will be chosen that lead to an improvement in storage size.

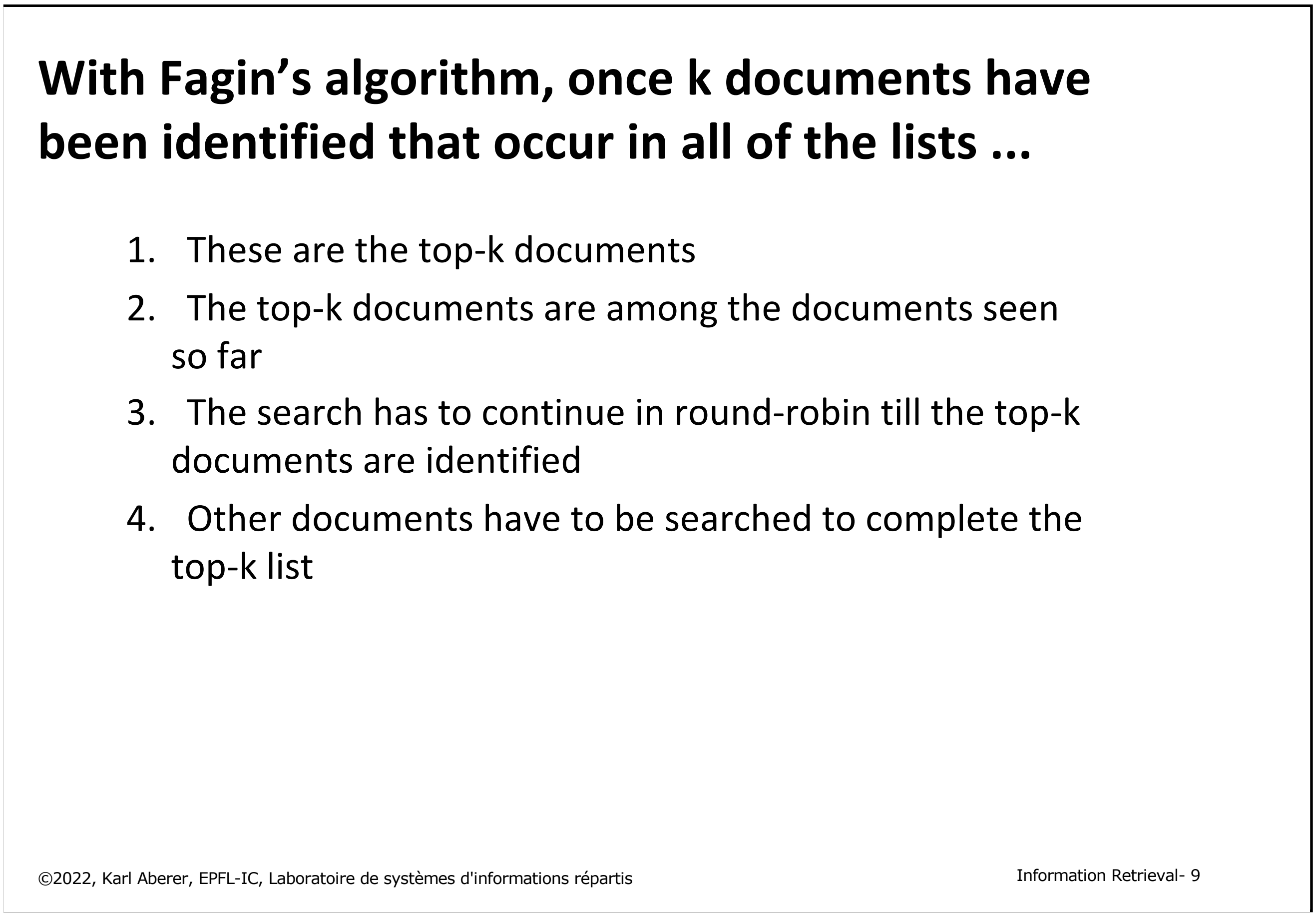


When applying Fagin’s algorithm for a query with three different terms for finding the k top documents, the algorithm will scan ...

1. 2 different lists
2. 3 different lists
3. k different lists
4. it depends how many rounds are taken

Answer 2

For answering a query with three terms, the posting lists of the three terms need to be inspected, therefore 3 different lists have to be scanned.

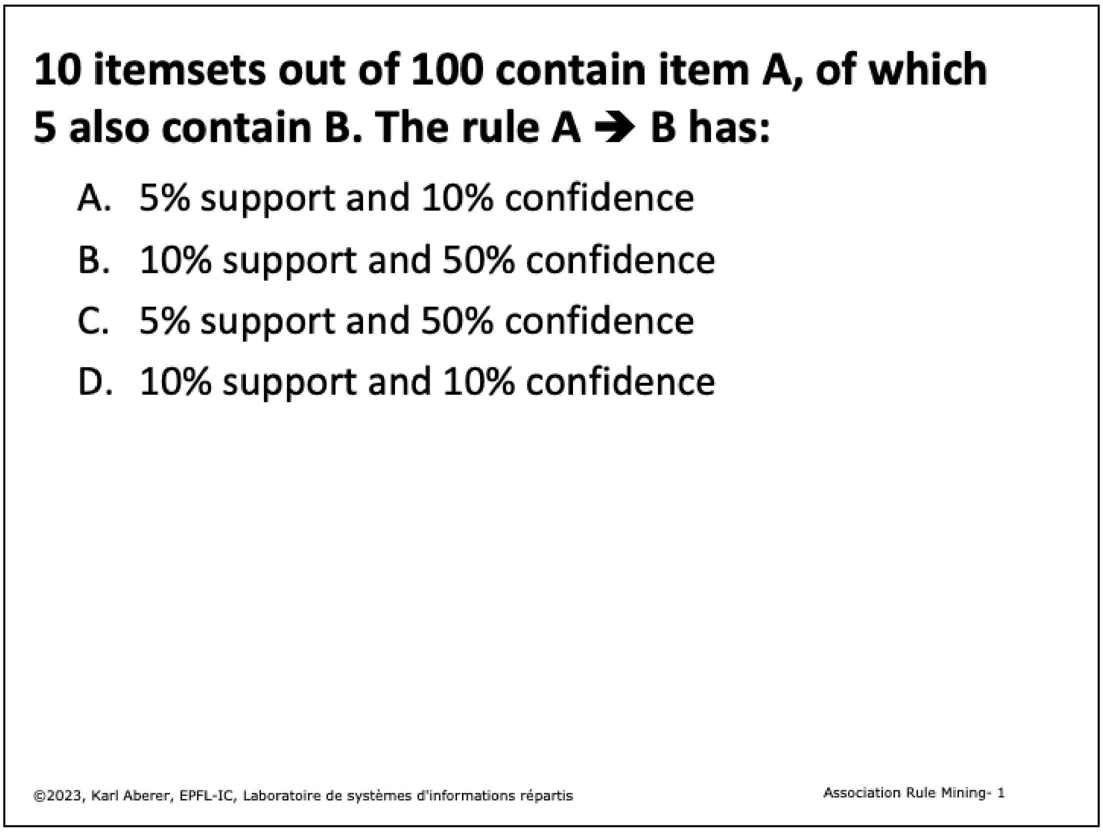


With Fagin’s algorithm, once k documents have been identified that occur in all of the lists ...

1. These are the top-k documents
2. The top-k documents are among the documents seen so far
3. The search has to continue in round-robin till the top-k documents are identified
4. Other documents have to be searched to complete the top-k list

Answer 2

In Fagin’s algorithm round-robin traversal of posting lists stops once k documents have been seen in all lists. At this point also other documents may have been seen. The top-k documents are among all the documents that have been seen, but not necessarily those that have been seen in all lists.



10 itemsets out of 100 contain item A, of which 5 also contain B. The rule A ➔ B has:

A. 5% support and 10% confidence
B. 10% support and 50% confidence
C. 5% support and 50% confidence
D. 10% support and 10% confidence

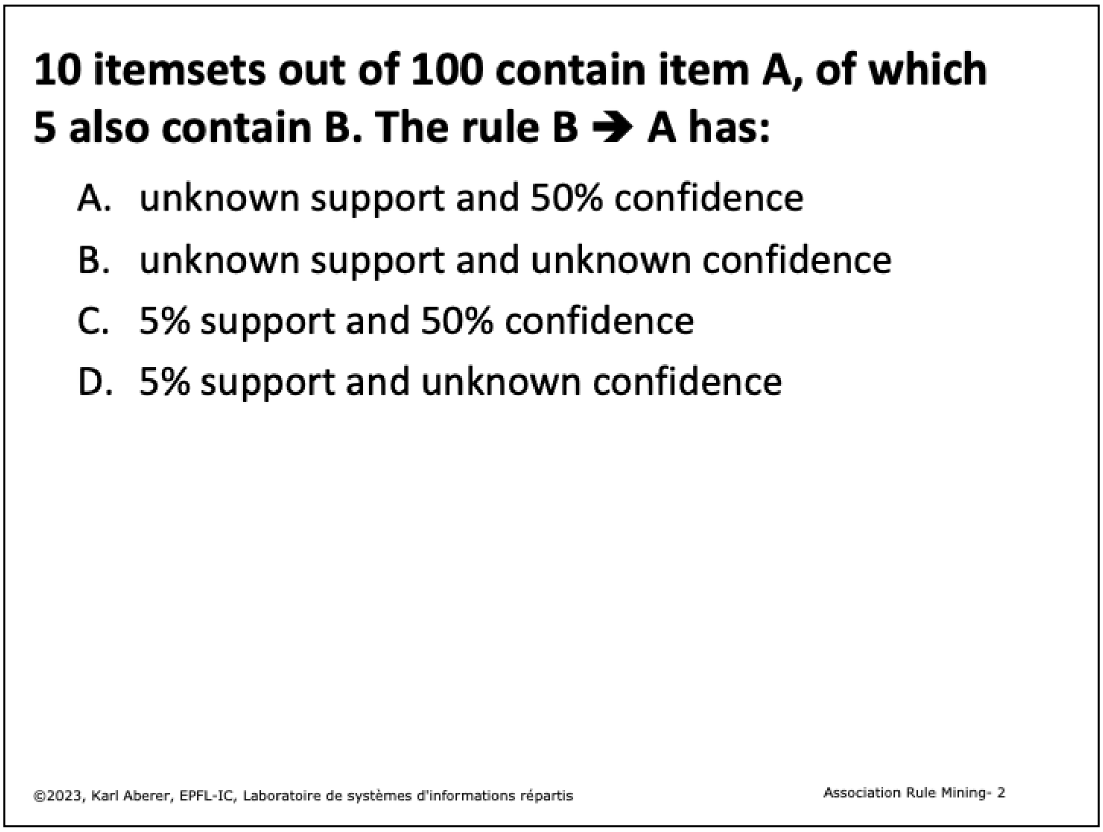
©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Association Rule Mining- 1

Answer C

The number of itemsets containing both A and B is 5, therefore the support is 5%.

In 5 out of 10 cases the rule is correct, therefore the confidence is 50%.



10 itemsets out of 100 contain item A, of which 5 also contain B. The rule B ➔ A has:

A. unknown support and 50% confidence
B. unknown support and unknown confidence
C. 5% support and 50% confidence
D. 5% support and unknown confidence

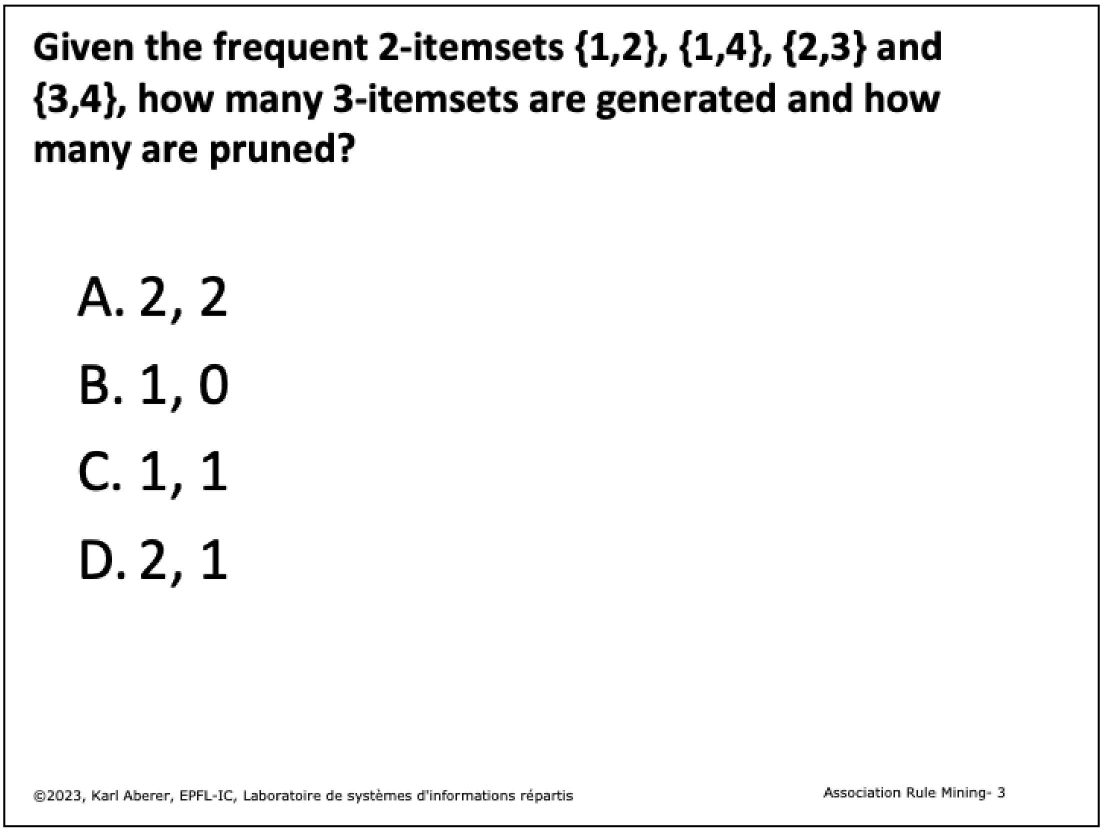
©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Association Rule Mining- 2

Answer D

As in the previous question the support is 5%.

Since we do not know the number of itemsets containing the item B, we cannot determing the confidence of the rule.



Given the frequent 2-itemsets {1,2}, {1,4}, {2,3} and {3,4}, how many 3-itemsets are generated and how many are pruned?

A. 2, 2

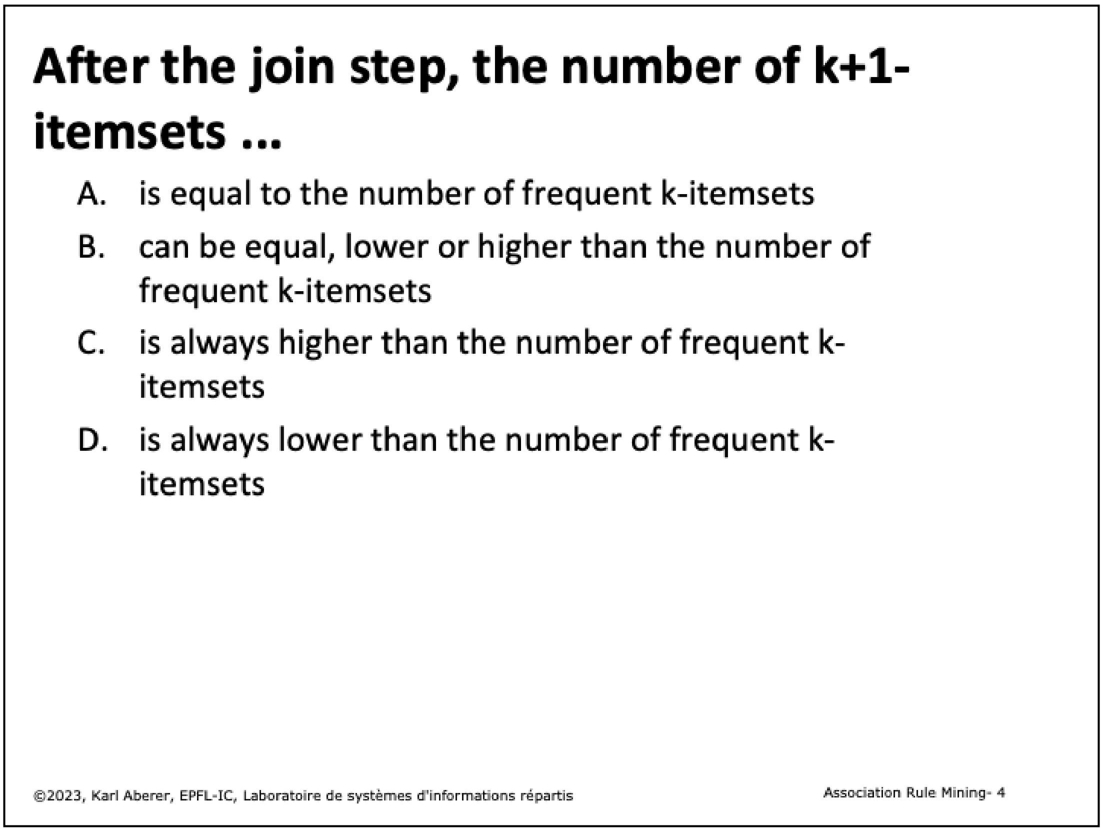
B. 1, 0

C. 1, 1

D. 2, 1

Answer C

Only one itemset will be generated, {1,2,4} for the first shared prefix {1}. Since the itemset {2,4} is not frequent, it will be pruned.



After the join step, the number of k+1-itemsets ...

A. is equal to the number of frequent k-itemsets

B. can be equal, lower or higher than the number of frequent k-itemsets

C. is always higher than the number of frequent k-itemsets

D. is always lower than the number of frequent k-itemsets

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Association Rule Mining- 4

Answer B

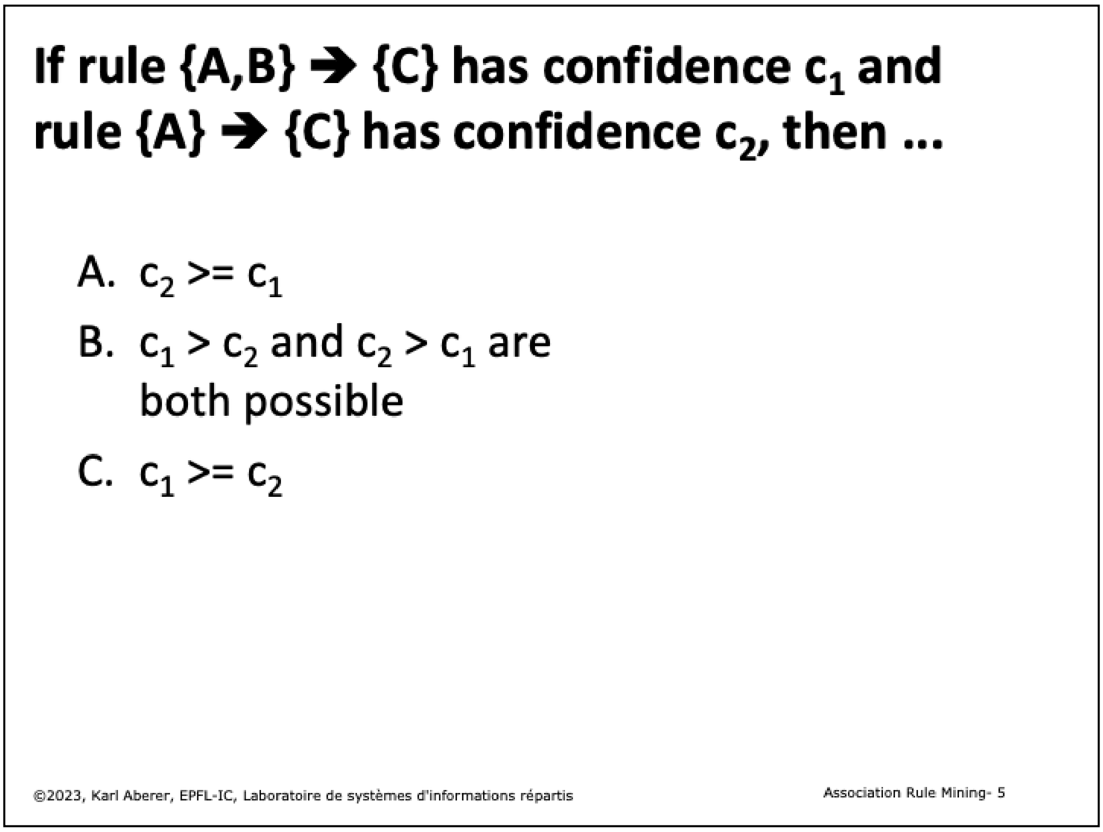
We cannot make any statement on the cardinality of the candidate set generated in the join step.

Examples:

{1,2} {1,3} will generate 1 3-itemset, so the cardinality is lower.

{1,2} {1,3} {1,4} will generate 3 3-itemsets, so the cardinality is equal.

{1,2} {1,3} {1,4} {1, 5} will generate 6 3-itemsets, so the cardinality is higher.



If rule {A,B} ➔ {C} has confidence c_1 and rule {A} ➔ {C} has confidence c_2, then …

A. $c_2 >= c_1$

B. $c_1 > c_2$ and $c_2 > c_1$ are both possible

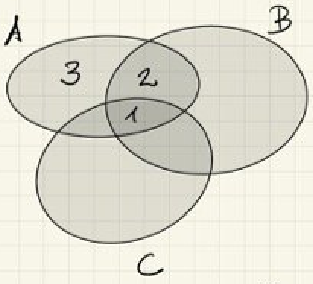
C. $c_1 >= c_2$

Association Rule Mining- 5

Answer B

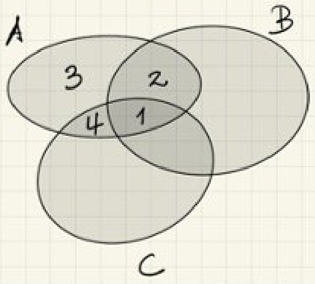
See example on next slide.

Example



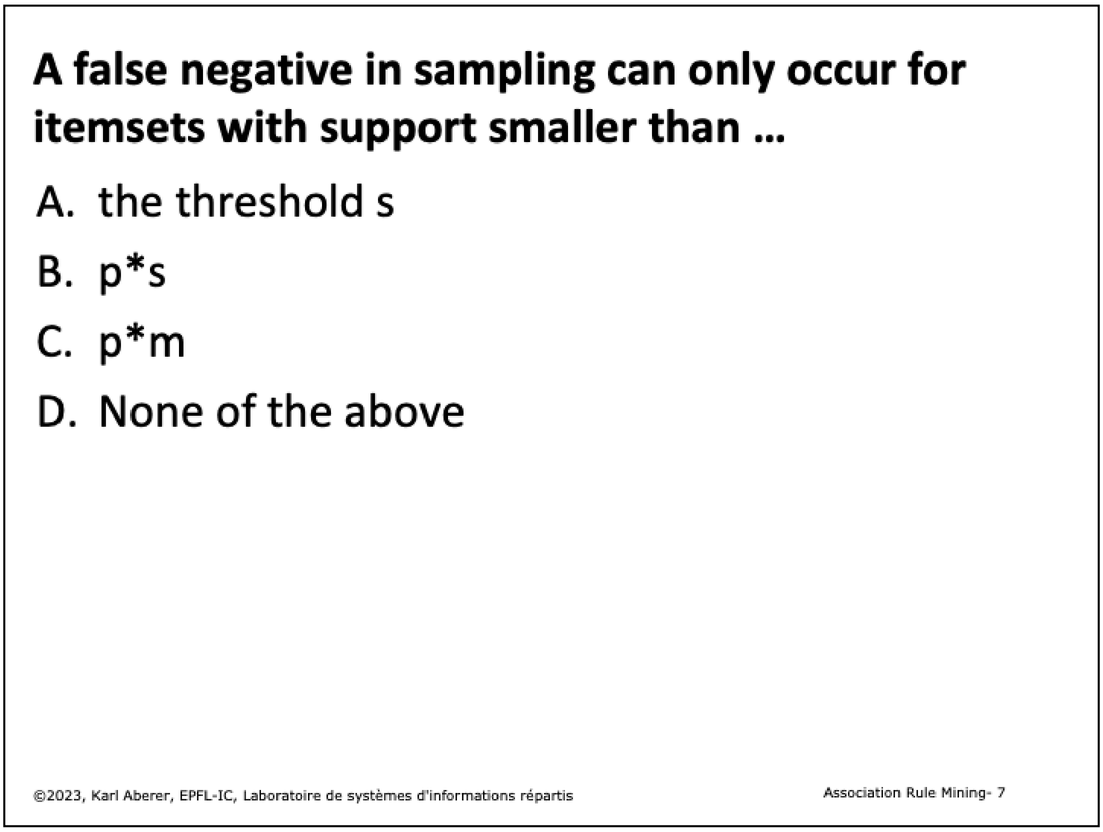
$$c_1 = \frac{s(A,B,C)}{s(A,B)} = \frac{1}{3}$$

$$c_2 = \frac{s(A,C)}{s(A)} = \frac{1}{6} = 0 \Rightarrow c_1 > c_2$$



$$c_1 = \frac{s(A,B,C)}{s(A,B)} = \frac{1}{3}$$

$$c_2 = \frac{S(A,C)}{s(A)} = \frac{5}{10} = \frac{1}{2} \Rightarrow c_1 < c_2$$



A false negative in sampling can only occur for itemsets with support smaller than …

A. the threshold s

B. p*s

C. p*m

D. None of the above

Answer D (under reasonable assumptions)

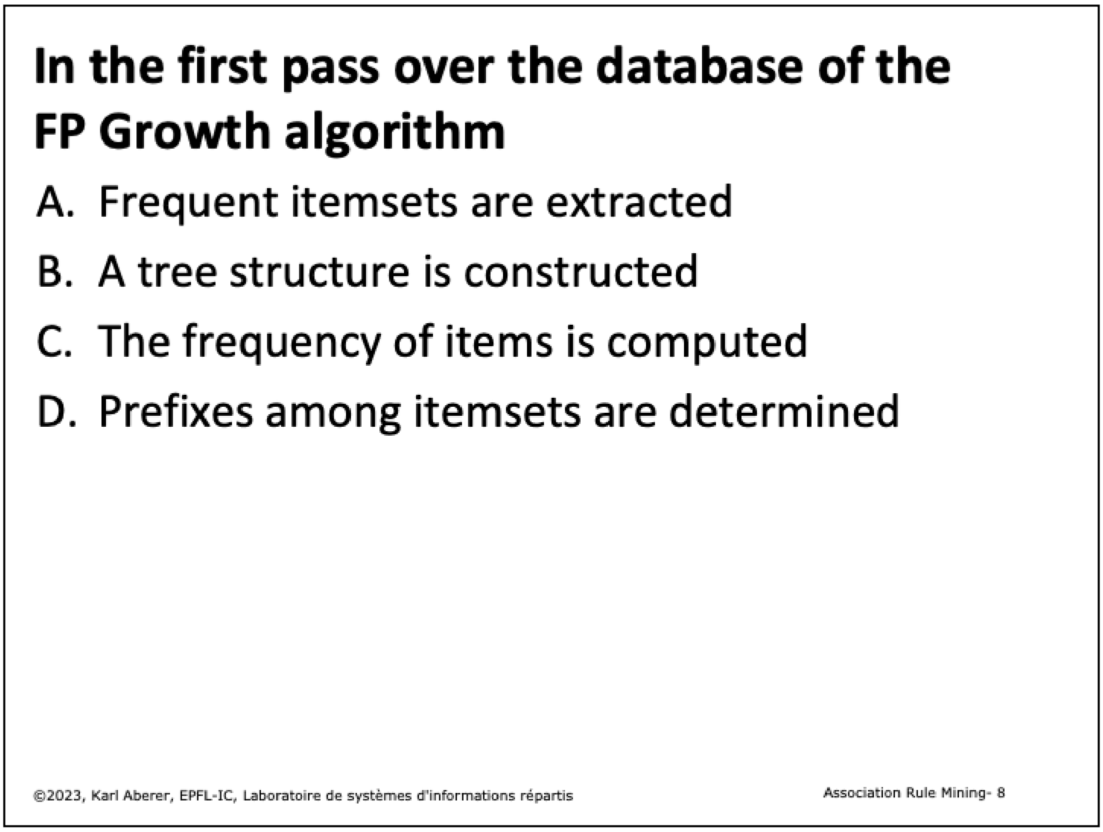
A false negative occurs if an itemset has a support above the threshold s, but in the sample less then p*s transactions are selected that contain the itemset (respectively less than the lowered threshold).

For each of the options in answers A, B, C it is easy to construct an adversarial example in which this is the case. Actually the more difficult case is when the itemset has the larger support count, as fewer remaining transactions remain to select from. Therefore, we conly consider the maximal possible size.

For answer A consider an itemset with support count larger than s. If m > (s – p*s+1) + p*m, then we can select the sample such that it contains at most p*s – 1 of the occurrences of the itemset. This is equivalent to saying that m > s, i.e. the database should have larger size than the support threshold. We never explicitly stated that this should be the case but is an obvious condition to satisfy.

For answer B the itemset has a support strictly smaller than s, and therefore for the same reasoning as in answer A, we can produce the counter-example.

For answer C we can select the occurrences without including the itemset from the (1-p)*m remaining occurrences if m > (m – p*m + 1) + p*m, which is equivalent to m > m-1. Here lowering the threshold could in the extreme case avoid having a false negative.



In the first pass over the database of the FP Growth algorithm

A. Frequent itemsets are extracted

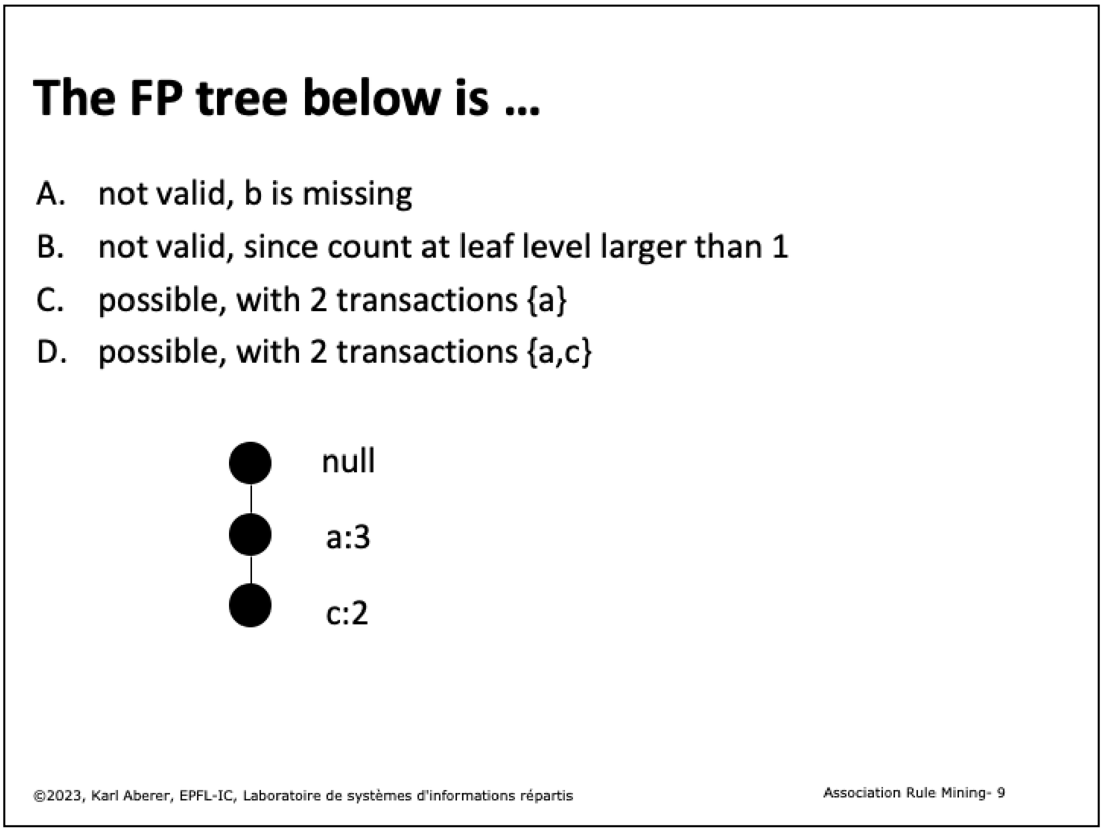
B. A tree structure is constructed

C. The frequency of items is computed

D. Prefixes among itemsets are determined

Answer C

The algorithms proceeds in two steps, where the first step has two passes over the database. The first pass is to compute the frequency of items, and the second to construct the tree structure. The frequent itemsets are extracted in the second step.



The FP tree below is …

A. not valid, b is missing

B. not valid, since count at leaf level larger than 1

C. possible, with 2 transactions {a}

D. possible, with 2 transactions {a,c}

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Association Rule Mining- 9

Answer D

Answer A can be excluded, since it is possible to have paths that are not including some elements. It simply means that the corresponding itemsets do not occur.

Answer B can be excluded since leaf counts can be larger than 1, of multiple itemsets corresponding to the path exist.

Answer C is not possible, since in the presence of two transactions with itemset {a} and the tree structure implying that there exist two transactions with itemset {a,c}, the minimal count of node a has to be 4.

Answer D is possible with three transactions {a}, {a,c}, {a,c}