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 $\beta > 0$
4. Yes, but only if $\gamma > 0$

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 $\beta > 0$. 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.