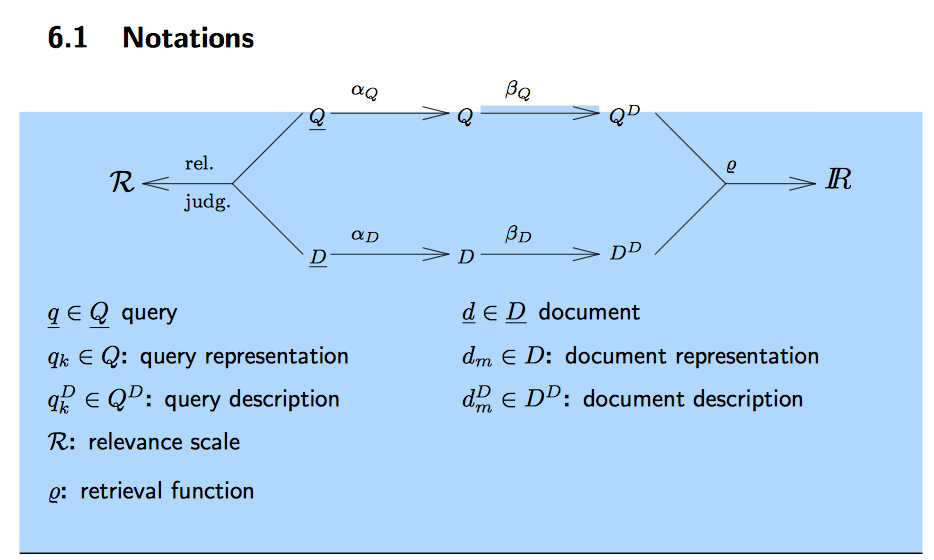
1. The BM25 retrieval model modifies the binary independence model by including term weighting. Explain what ‘binary independence’ means and what the ranking function for the binary independence model would be.
2. The BM25 *document* term weighting component includes three parameters: *k* 1  , *b* , and *K* , which is a function of the other two. (The *k* 2  parameter is used for query term weighting.) Explain in words the impact of varying the values of these parameters, and submit a graph of the term-weighting component of BM25 as a function of term frequency for *k* 1 =1 and *K*=1 .

ANSWER1:

The Binary Independence Model (BIM) is a probabilistic information retrieval technique that makes some simple assumptions to make the estimation of document/query similarity probability feasible.



BIM assigns weights to query terms.

ASSUMPTIONS:

There is an ideal answer set (relevant documents) for a given user query.

• We do not know the description of the ideal set (its properties).

• We have index terms that have semantics that can characterize the properties of the ideal answer set.

• We make an initial guess about the ideal set at query time.

• We give the answer to the user and “hopefully” we get some feedback that would allow us to further refine the description of the ideal set.

* The probability P(R|d,q) that a document is relevant derives from the probability of relevance of the terms vector of that document P(R|x,q). By using the Bayes rule we get:

P(R|x,q) = \frac{P(x|R,q)\*P(R|q)}{P(x|q)}

Where,

P(x|R=1,q) and P(x|R=0,q) are the probabilities of retrieving a relevant or nonrelevant document, respectively.

* Since a document is either relevant or nonrelevant to a query we have that:

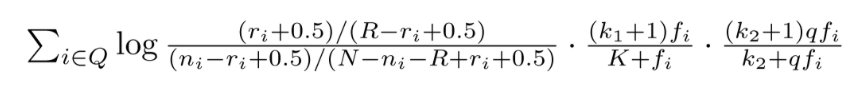
P(R=1|x,q) + P(R=0|x,q) = 1

IMPORTANT POINTS:

* Basically every document is represented by terms. We relate all term occurrences in the document.
* **The term Binary means Boolean** : Where documents are represented as binary vectors of terms.

**The term Independence** refers to terms occurring in the document independently.

* Terms which do not appear in a query do not affect the overall outcome.
* Term occurrence in queries is conditionally independent and relevance of documents depends only on the terms present in the query.
* Formula used:



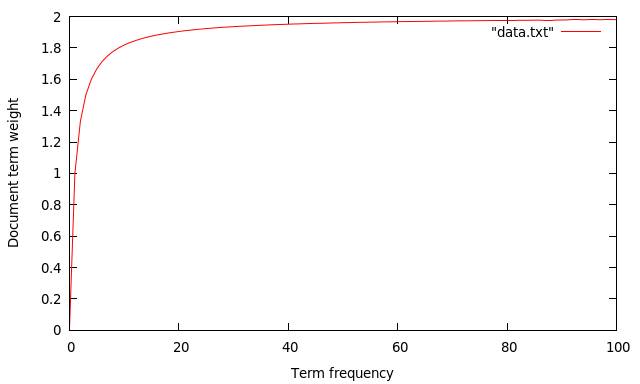
￼– K = K1((1 – b) + b. dl/avdl)

where : k1, k2 and K are parameters whose values are set

empirically

* dl is doc length
* avdl is average document length
* We conclude that a weight assignment to query terms yields better retrieval effectiveness than if query terms are equally weighted

ANSWER 2:



X-axis represents : Term frequency

Y-aaxis represents: Document term weight.

K2 is used for query term weighting .

K = K1((1 – b) + b. dl/avdl)

A document may be assigned with varying scores by different search engines for the same input query, although the document contains exactly the same amount of relevant information.

Assuming that every retrieved document can be classified as relevant and non- relevant according to how much it meets user’s information need, the ranked list of documents is in fact a mixture of both relevant and non-relevant documents.

The inference from the graphs is that there is no impact of varying the values of the given paramaters because K is independent of the query terms. Increment in value of K will be because more relevant documents are retrieved. When more query terms will be added the Standard deviation becomes smaller therefore an exponential shaped distribution should appear for non-relevant documents.

(Please refer to the graphs attached and the one above)