PROBALISTIC INFORMATION RETRIEVAL

- Users start with information needs, which they translate into query representations.
- Similarly, there are documents which are converted into document representations.
- Based on these two representations, a system tries to determine how well documents satisfy information needs.
- In the Boolean or vector space model, matching is done with index terms.
- Given the query and document representations, a system has an uncertain guess of whether a document has content relevant to the information need.
- Probability theory provides a principle foundation of such reasoning under uncertainty.

THE 1/0 LOSS CASE

- The user issues a query and an ordered list of documents is returned.
- For a query 'q' and a document 'd' in the collection, let $R_{d,\ q}$ be an indicator random variable that says whether 'd' is relevant with respect to a given query 'q'.
- That is, it takes on a value of 1 when the document is relevant and 0 otherwise.
- Using a probabilistic model, the order in which to present documents to the user is to rank documents by their estimated probability of relevance with respect to information need, $P(R=1|d,q)$, which is the basis of Probability Ranking Principle (PRP).
- Ranking of the documents in the collection is in order of decreasing probability of relevance to the user who submitted the request.
- You lose a point for either returning a non-relevant document or failing to return a relevant document.
- Such a binary situation where you are evaluated on your accuracy is called 1/0 loss.
- PRP rank all the documents in the decreasing order of $P(R=1|d,q)$
- d is relevant iff,

$$P(R=1|d,\ q) > P(R=0|d,\ q).$$

BINARY INDEPENDENCE MODEL (BIM)

- Binary is equivalent to Boolean.
- Documents and queries are both represented as binary term incidence vectors.
- i.e. a document d is represented by the vector $x(x_1, x_2, \cdots\cdots\cdots\cdots, x_m)$, where $x_t = 1$ if term t is present in document d and $x_t = 0$ if t is not present in d.
- Similarly, query q is represented by query vector q.
- Independence means that terms are modeled as occurring in documents independently.
- i.e. the model recognizes no association between terms.

$$P(R=1|x,q) = \frac{P(x|R=1,q)\,P(R=1|q)}{P(x|q)}$$

$$P(R=0|x,q) = \frac{P(x|R=0,q)\,P(R=0|q)}{P(x|q)}$$

- But some assumptions like terms that are independent of BIM can be removed.
- Example: term pairs "Hong" and "Kong"are strongly dependent.
- Others are Stock, Exchange, New, York, etc.